

Automated COVID-19 Grading with Convolutional Neural Networks in Computed Tomography Scans: A Systematic Comparison

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

Abstract—Amidst the ongoing pandemic, the assessment of computed tomography (CT) images for COVID-19 presence can exceed the workload capacity of radiologists. Several studies addressed this issue by automating COVID-19 classification and grading from CT images with convolutional neural networks (CNNs). Many of these studies reported initial results of algorithms that were assembled from commonly used components. However, the choice of the components of these algorithms was often pragmatic rather than systematic and systems were not compared to each other across papers in a fair manner. We systematically investigated the effectiveness of using 3D CNNs instead of 2D CNNs for seven commonly used architectures, including DenseNet, Inception, and ResNet variants. For the architecture that performed best, we furthermore investigated the effect of initializing the network with pre-trained weights, providing automatically computed lesion maps as additional network input, and predicting a continuous instead of a categorical output. A 3D DenseNet-201 with these components achieved an area under the receiver operating characteristic curve (AUC) of 0.930 on our test set of 105 CT scans and an AUC of 0.919 on a publicly available set of 742 CT scans, a substantial improvement in comparison with a previously published 2D CNN. This paper provides insights into the performance benefits of various components for COVID-19 classification and grading systems. We have created a challenge on grand-challenge.org to allow for a fair comparison between the results of this and future research.

Impact Statement—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.

Index Terms—3D convolutional neural network, CO-RADS, COVID-19, deep learning, medical imaging.

I. 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 [1]. 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 [2]–[23]. Automatic COVID-19 classification methods have already been deployed in several medical centers [8].

By far the most common technique for automatic COVID-19 classification from CT images is the Convolutional Neural Network (CNN) [24], [25], which is the current state-of-the-art for image classification [26]. The works that use this approach can be divided into those that use 2D CNNs [2], [6], [7], [11], [13], [15], [18]–[20], [22] and those that use 3D CNNs [4], [9], [10], [12]–[14], [16], [17], [23]. 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 [27]. 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 [28] 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

Manuscript submitted September 25, 2020; revised March 3, 2021; revised April 28, 2021; revised June 2, 2021; revised July 28, 2021. (Coen de Vente and Luuk H. Boulogne contributed equally to this work.) (Corresponding author: Luuk H. Boulogne, e-mail: luuk.boulogne@radboudumc.nl)

Coen de Vente is with the Radboud University Medical Center, Donders Institute for Brain, Cognition and Behaviour, Department of Medical Imaging, Nijmegen, The Netherlands. Luuk H. Boulogne, Kiran Vaidhya Venkadesh, Cheryl Sital, Nikolas Lessmann, Colin Jacobs and Bram van Ginneken are with the Radboud University Medical Center, Radboud Institute for Health Sciences, Department of Medical Imaging, Nijmegen, The Netherlands. Clara I. Sánchez and Coen de Vente are with the Informatics Institute, Faculty of Science, University of Amsterdam, The Netherlands.

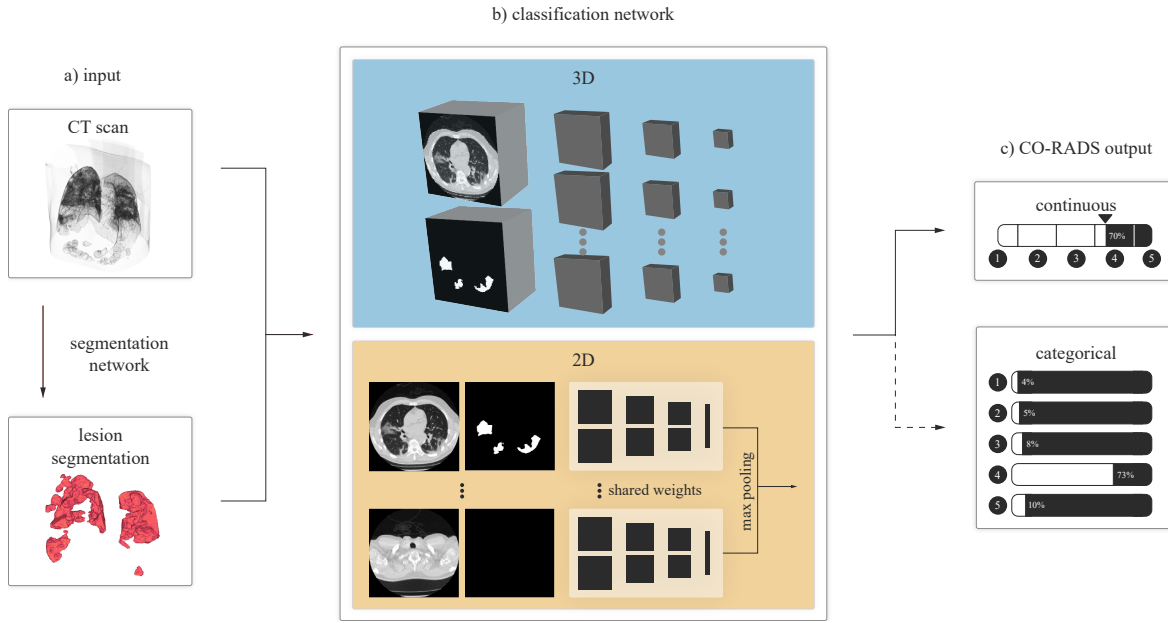


Fig. 1. Schematic representation of the different components used for CO-RADS grading from CT scans using convolutional neural networks in patients suspected with COVID-19. This processing pipeline was used in all experiments of this work. (a) The input CT scan is fed into a lesion segmentation network. The CT and the lesion segmentation are used as separate input channels to the classification network as described in Section III-C1. 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 III-C2. 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 III-C4 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.

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 [29], [30].

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 [4], [14]. Also, the same validation methods, such as receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC), have been reported across different studies [2], [4], [6]–[10], [12]–[15], [18]–[20], [22], [23]. 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 [31], 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 [32] for eval-

uating 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.

II. BACKGROUND

3D CNNs were initially proposed for processing video data [25], 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 expanding the 2D filters into 3D [33]. 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 [34], [35].

The large majority of the architectures used for COVID-19 classification from CT scans in previous works [2], [4]–[10], [12], [14]–[19], [19], [20], [22], [23], [36] are heavily or completely based on the ResNet [37], DenseNet [38], or Inception [39] architecture families. Especially ResNet architectures have been used frequently [2], [6], [8]–[10], [15]–[20], [36]. Some works did not use a full ResNet architecture, but did incorporate residual blocks into their model [5], [22]. Architectures from the DenseNet [4], [19], [23] and Inception [7], [14] families have been used less frequently. Other architectures such as VGG-19 [40], Inception-ResNet-v2 [41], NASNet [42], and EfficientNet [43] have also been

used in research for COVID-19 classification from CT scans [36], [44]–[47]. 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 [48] and COVID-19 classification specifically [49]. 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 [50] 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 [48]. It has been used in several 2D CNN COVID-19 classification methods [2], [6], [7], [18], [20]. Pre-trained weights have also been used for 3D CNN-based methods. Wang *et al.* [4] 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 [33] and chest CT classification [34] 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 [4], [6], [14], [16], [18] and/or by suppressing non-lung tissue [2], [4], [6], [8]–[10], [12], [15], [17]. Yang *et al.* [19] 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.* [14] 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 [49]. Previous works have trained models to discern between COVID-19 positive and negative patients [3], [5], [6], [12], [15], [16], [18]–[20], [22], [23], COVID-19 positive patients and patients with other types of pneumonia [4], [7], [9], and between all three [2], [10], [17]. In this work, we followed Lessmann *et al.* [14] and trained our models to produce CO-RADS [27] 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 [27]. Scoring systems, like CO-RADS, have been advocated for better communication between radiologists and other healthcare providers [14], [27].

III. METHODOLOGY

A. Data

1) *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

TABLE I
NUMBER OF CT IMAGES IN INTERNAL DATASET.

	CO-RADS					Total	Neg	Pos
	1	2	3	4	5			
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 II
NUMBER OF CT IMAGES IN EXTERNAL DATASET.

Grade [13]					Total	Neg	Pos
Control	Mild	Regular	Severe	Critically ill			
207	23	363	117	32	742	207	535

of Radboudumc (CMO2016-3045, Project 20027) prior to the study. Further details such as imaging parameters can be found elsewhere [14].

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 I. All data splits were made such that all scans from a patient with multiple visits ended up in the same split.

2) *External test data*: For external evaluation, we used the publicly available CT images and clinical features for COVID-19 dataset (iCTCF) dataset [13], [31]. 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 [13], we did not use them for testing our models. The distribution of the other classes is displayed in Table II. The grading system uses etiological laboratory confirmation and other factors such as clinical features and CT imaging [13]. 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 [31].

B. 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 II 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.* [2]. 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.

C. Ablation study

We investigated whether additional model components had an effect on COVID-19 classification performance in an ablation study. Fig. 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).

1) *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 [29] trained by Lessmann *et al.* [14], which segments ground-glass opacities (GGOs) and consolidations, provided the lesion segmentations. GGOs and consolidations are biomarkers with major importance in diagnosing COVID-19 [27].

2) *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.

3) *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.

4) *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.* [51] explored this approach for prostate cancer grading and found that it outperformed other regression and categorical output methods.

D. 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 mm³ 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 [52]. 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 × 240 pixels around the center of the mask. Following previous research with I3D models [33]–[35], 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.

E. 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 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 10000 training batches and the QWK on the validation set

for the stopping criterion. Gradient checkpointing [53] 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.

F. 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 IV were obtained from ensembles unless stated otherwise.

G. 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 [2], [4], [6]–[10], [12]–[14], diagnostic performance was evaluated using the AUC and ROC curves.

We calculated 95% confidence intervals (CIs) with non-parametric bootstrapping and 1000 iterations [54]. Statistical significance was computed with the same bootstrapping method [55].

The AUCs that our models achieved on the external test set are additionally listed on the Grand Challenge platform [32] 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.

IV. RESULTS

A. Architecture selection

Fig. 2 shows the QWK and AUC for the different 2D and 3D architectures. Table III 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

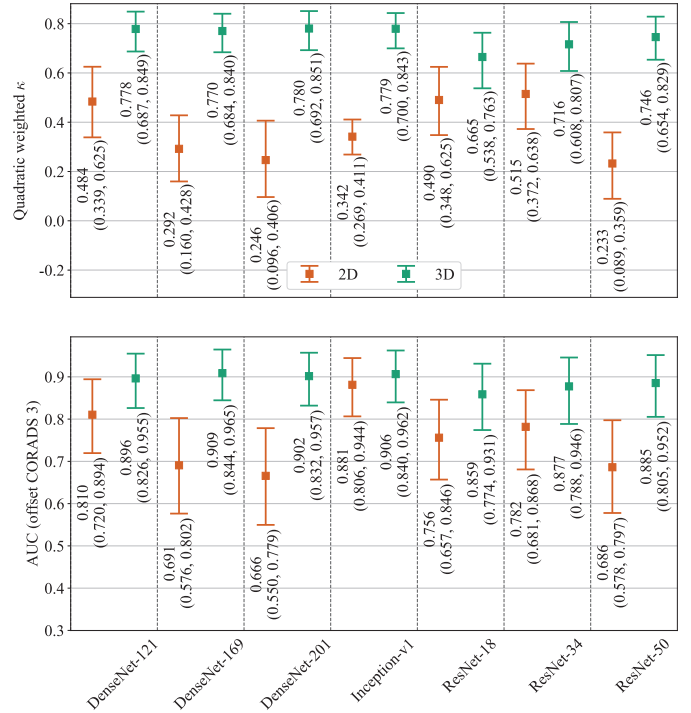


Fig. 2. Performance of 2D and 3D CNN architectures on the internal test set for the task of CO-RADS grading from CT images is shown in QWK and AUC, respectively. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans) and CO-RADS 3-5 as the positive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

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

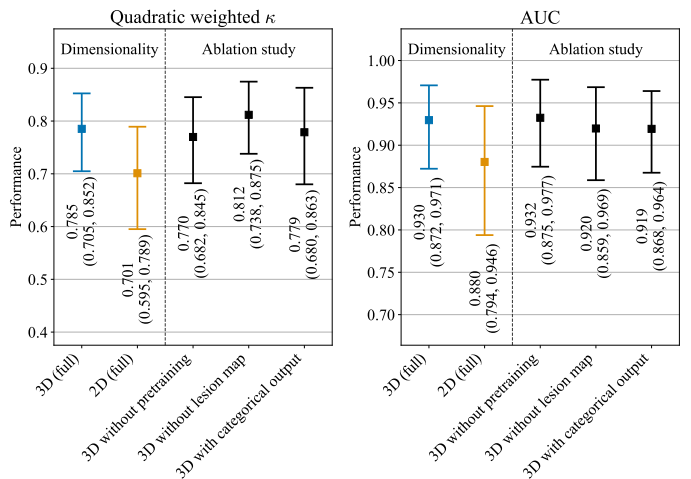


Fig. 3. Comparison of 2D and 3D DenseNet-201 models and ablation study with this architecture for the task of CO-RADS grading from CT images. The analysis was performed on the internal test set. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans)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

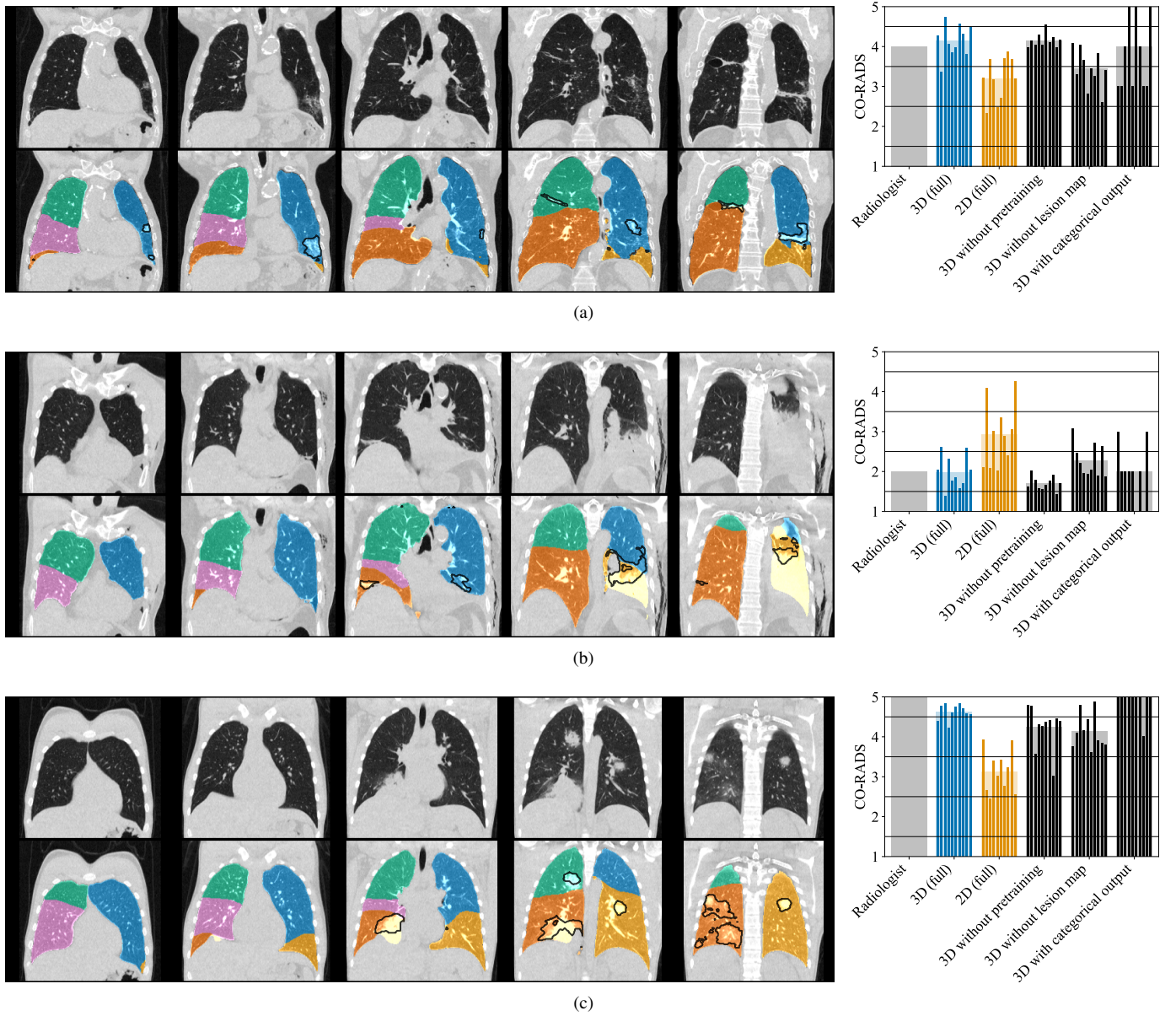


Fig. 4. Example input-output pairs for the task of CO-RADS grading on the internal test set for the trained DenseNet-201 ensembles. Input examples are shown on the left. Top row: Coronal slices of an input CT scan. Bottom row: Lung segmentation used for centering and cropping are displayed with colored 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 intrapulmonary 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."

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.

B. 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 and 6 show the corresponding CIs and ROC analyses respectively. Fig. 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.* [2], 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. 5 shows confusion matrices for the two dimensionali-

TABLE III
ARCHITECTURE PROPERTIES

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

ties. 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.

C. 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. 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 performance 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,

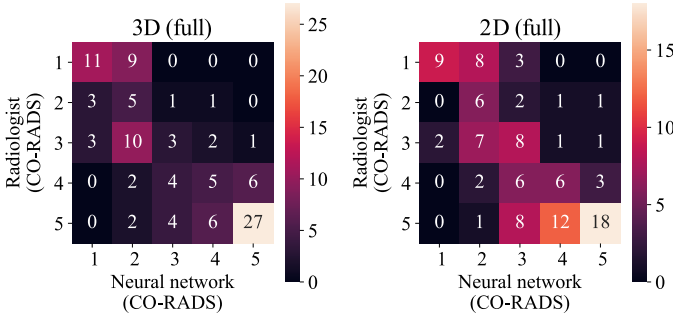


Fig. 5. Confusion matrices for CO-RADS grading of the 2D and 3D DenseNet-201 model predictions on the internal test set. These models were trained with transfer learning, lesion maps and produced continuous output. The true label reference is from the radiology report. Cells contain the number of CT scans.

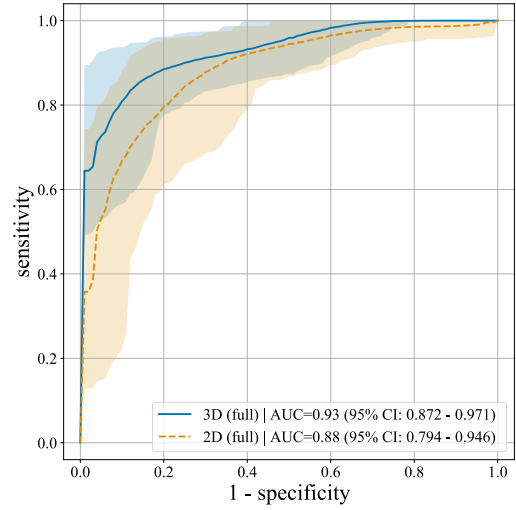


Fig. 6. ROC analysis for the 2D and 3D DenseNet-201 models on the internal test set from Radboudumc (105 CT scans) for the task of CO-RADS grading. The analysis was performed with CO-RADS 1 and 2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans). It was performed for the full 2D and 3D models trained with transfer learning, lesion maps and continuous output.

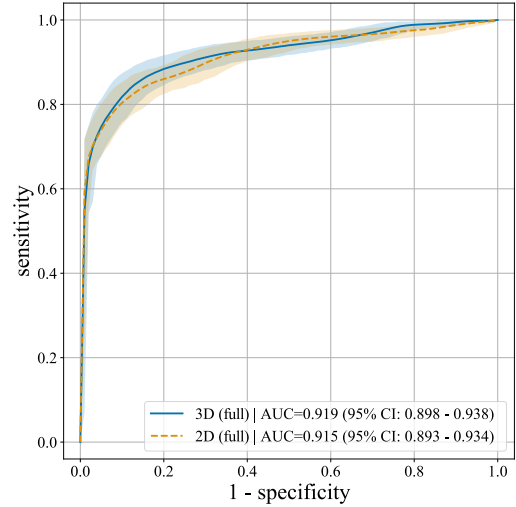


Fig. 7. ROC analysis for the 2D and 3D DenseNet-201 models on the external iCTCF test set (742 CT scans) for the task of COVID-19 classification. The analysis was performed with 207 COVID-19 negative (Control) cases and 535 positive (Mild, Regular, Severe, Critically ill) cases.

$p = .324$). Fig. 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 lesion input required more iterations (32 750).

D. External evaluation

Fig. 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, $p = .215$).

E. Lesion segmentation model

For a single patch the lesion segmentation model inference time was $178.66 \text{ ms} \pm 14.56 \text{ ms}$, 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.* [14].

V. 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 [2], 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.* [13] 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 [28], 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 [27].

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 obtained 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.

VI. 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 [29], [30]. 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 online Grand Challenge platform [32]. 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.

REFERENCES

- [1] 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)," *European Radiology*, vol. 30, no. 9, pp. 1–9, 2020.
- [2] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, no. 2, pp. 65–72, 2020.
- [3] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," *arXiv:2003.09424*, 2020.
- [4] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu *et al.*, "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis," *European Respiratory Journal*, vol. 56, no. 2, 2020.
- [5] 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," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 39, no. 7, pp. 1–11, 2020.
- [6] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Zha *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, early access.
- [7] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using CT images to screen for corona virus disease (covid-19)," *European radiology*, vol. 31, no. 8, pp. 1–9, 2021.
- [8] 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," *Applied Soft Computing*, vol. 98, p. 106897, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494620308358>
- [9] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song *et al.*, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, 2020.
- [10] 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," *IEEE Transactions on Medical Imaging*, 2020.
- [11] 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," *Scientific Reports*, vol. 10, no. 1, p. 19196, Nov 2020. [Online]. Available: <https://doi.org/10.1038/s41598-020-76282-0>
- [12] 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," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.
- [13] W. Ning, S. Lei, J. Yang, Y. Cao, P. Jiang, Q. Yang, J. Zhang, X. Wang, F. Chen, Z. Geng *et al.*, "iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia," *Research Square*, 2020.
- [14] 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," *Radiology*, 2020.
- [15] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung *et al.*, "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.
- [16] Y. Li, D. Wei, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian *et al.*, "Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2787–2797, 2020.
- [17] 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," *Cancer Management and Research*, vol. 12, p. 2979, 2020.
- [18] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su *et al.*, "Rapid ai development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.
- [19] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:2003.13865*, 2020.
- [20] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan, "Coronavirus detection and analysis on chest CT with deep learning," *arXiv preprint arXiv:2004.02640*, 2020.

- [21] D. Di, F. Shi, F. Yan, L. Xia, Z. Mo, Z. Ding, F. Shan, B. Song, S. Li, Y. Wei *et al.*, "Hypergraph learning for identification of COVID-19 with CT imaging," *Medical Image Analysis*, vol. 68, p. 101910, 2020.
- [22] X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi, "COVID-AL: The diagnosis of COVID-19 with deep active learning," *Medical Image Analysis*, vol. 68, p. 101913, 2020.
- [23] S. A. Harmon, T. H. Sanford, S. Xu, E. B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou *et al.*, "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nature communications*, vol. 11, no. 1, pp. 1–7, 2020.
- [24] 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.
- [25] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, 2013.
- [26] 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.
- [27] M. Prokop, V. 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," *Radiology*, p. 201473, 4 2020.
- [28] 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. Bellotti, 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," *Medical Image Analysis*, vol. 14, pp. 707–722, 12 2010.
- [29] 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," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [30] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2065–2081, 2020.
- [31] 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.
- [32] C. de Vente and L. H. Boulogne, "Grand Challenge - COVID-19 CT Classification challenge," <https://covid19.grand-challenge.org/>, accessed: 2020-09-01.
- [33] 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.
- [34] 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," *Nature medicine*, vol. 25, pp. 954–961, 2019.
- [35] I. W. Harsono, S. Liawati, and T. W. Cenggoro, "Lung nodule detection and classification from thorax CT-scan using RetinaNet with transfer learning," *Journal of King Saud University-Computer and Information Sciences*, 2020, to be published.
- [36] 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," *Pattern Recognition*, vol. 114, p. 107848, 2021.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [41] 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.
- [42] 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.
- [43] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.
- [44] 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 *et al.*, "Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT," *Radiology*, vol. 296, no. 3, pp. E156–E165, 2020.
- [45] S. A. A. Ahmed, M. C. Yavuz, M. U. Sen, F. Gulsen, O. Tutar, B. Korkmaz, C. Samanci, S. Sirolu, R. Hamid, A. E. Eryurekli *et al.*, "COVID-19 detection in computed tomography images with 2D and 3D approaches," *arXiv preprint arXiv:2105.08506*, 2021.
- [46] P. K. Chaudhary and R. B. Pachori, "FBSED based automatic diagnosis of COVID-19 using X-ray and CT images," *Computers in biology and medicine*, vol. 134, p. 104454, 2021.
- [47] 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," *Medical Physics*, vol. 48, no. 8, pp. 4334–4349, 2021.
- [48] 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.
- [49] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare *et al.*, "Automated detection and forecasting of COVID-19 using deep learning techniques: A review," *arXiv preprint arXiv:2007.10785*, 2020.
- [50] 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.
- [51] 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," *IEEE Transactions on Biomedical Engineering*, 2020.
- [52] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2664–2675, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07443>
- [53] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.
- [54] C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," *Academic Radiology*, vol. 7, pp. 413–419, 2000.
- [55] 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.



of eye diseases from retinal imaging.

Coen de Vente obtained his B.S. and M.S. degrees in biomedical engineering at the Eindhoven University of Technology, the Netherlands, in 2017 and 2019, respectively. There, he followed a track on medical imaging, a collaborative effort with the University Medical Center Utrecht, the Netherlands.

Since 2019, he is a Ph.D. candidate at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands. His current research focuses on deep learning techniques for medical image analysis and screening



Luuk H. Boulogne received his B.S. and M.S. degrees in artificial intelligence at the University of Groningen, the Netherlands, in 2016 and 2018, respectively.

Since 2019, he is a Ph.D. candidate at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands. His research focuses on predicting the effects of lung volume reduction surgery on a patient's lung function.



Kiran Vaidhya Venkadesh completed his B.Tech and M.Tech in the Engineering Design department at the Indian Institute of Technology, Madras in 2016, with a specialization in Biomedical Design.

After his graduation, he worked at Predible Health and developed deep learning solutions for medical image analysis for the Indian healthcare system. Since 2019, he is a Ph.D. candidate at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands.

There, he is working on early lung cancer detection with an emphasis on temporal analysis on chest CT scans using deep learning.



Cheryl Sital obtained her B.S. and M.S. degrees in biomedical engineering at the Eindhoven University of Technology, the Netherlands, in 2017 and 2019, respectively. There, she followed a track focused on medical imaging, a collaborative effort with the University Medical Center Utrecht, the Netherlands.

In 2020, she joined the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands as a Ph.D. candidate. She works on deep learning techniques for improved assessment of oncological CT scans.



Nikolas Lessmann obtained his B.S. and M.S. degrees in biomedical engineering at the University of Lübeck, Germany, in 2009 and 2013, respectively. He obtained his Ph.D. degree on the subject of Medical Image Analysis at the University Medical Center Utrecht, the Netherlands.

Since 2019, he is a tenure-track researcher in the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands, where he is leading the research group on musculoskeletal image analysis with machine

learning and artificial intelligence.



Colin Jacobs received his B.S. and M.S. degrees in biomedical engineering at the Eindhoven University of Technology, the Netherlands, in 2008 and 2010, respectively. He obtained his Ph.D. on the subject of Medical Image Analysis at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands. His Ph.D. research focused on the automatic detection and characterization of pulmonary nodules in thoracic CT scans.

Since 2017, he is Assistant Professor at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands. There, he leads the research line on lung cancer image analysis. From 2010 to 2013 he worked as a Biomedical Engineer for Fraunhofer MEVIS, Germany.



Clara I. Sánchez graduated in telecommunication engineering at the University of Valladolid, Spain, in 2003. She received her Ph.D. on the subject of Medical Image Analysis at the University of Valladolid, Spain in 2008.

From 2008 to 2010, she worked as a postdoctoral researcher at the University Medical Center Utrecht, the Netherlands. From 2010 until 2020 she worked at the Diagnostic Image Analysis Group, Department of Medical Imaging, Radboudumc, Nijmegen, the Netherlands. There she led the research line focused

on Computer-Aided Diagnosis of Retinal Images. Since 2020, she is Full Professor at the University of Amsterdam.



Bram van Ginneken studied physics at the Eindhoven University of Technology, the Netherlands and received his M.S. in physics at the Utrecht University, the Netherlands, in 1995. In 2001 he obtained his Ph.D. at the Utrecht University on the subject of Medical Image Analysis, the Netherlands.

Since 2012, he is Full Professor of Medical Image Analysis at Radboudumc and chairs the Diagnostic Image Analysis Group. He also works for Fraunhofer MEVIS in Bremen, Germany, and is a founder of Thirona, a company that develops software and

provides services for medical image analysis. He is member of the Editorial Board of Medical Image Analysis. He pioneered the concept of challenges in medical image analysis.