Medical Image Analysis 97 (2024) 103230

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

The STOIC2021 COVID-19 AI challenge: Applying reusable training methodologies to private data

Luuk H. Boulogne [a,*,1], Julian Lorenz [b,2], Daniel Kienzle [b,2], Robin Schön [b,2], Katja Ludwig [b,2], Rainer Lienhart [b,2], Simon Jégou [c,3], Guang Li [d,4], Cong Chen [d,4], Qi Wang [d,4], Derik Shi [d,4], Mayug Maniparambil [e,5], Dominik Müller [f,b,6], Silvan Mertes [f,6], Niklas Schröter [f,6], Fabio Hellmann [f,6], Miriam Elia [f,6], Ine Dirks [g,n,7], Matías Nicolás Bossa [g,n,7], Abel Díaz Berenguer [g,n,7], Tanmoy Mukherjee [g,n,7], Jef Vandemeulebroucke [g,n,7], Hichem Sahli [g,n,7], Nikos Deligiannis [g,n,7], Panagiotis Gonidakis [g,n,7], Ngoc Dung Huynh [h,8], Imran Razzak [i,8], Reda Bouadjenek [h,8], Mario Verdicchio [j,9], Pasquale Borrelli [j,9], Marco Aiello [j,9], James A. Meakin [a,1], Alexander Lemm [k,1], Christoph Russ [k,1], Razvan Ionasec [k,1], Nikos Paragios [l,d,1], Bram van Ginneken [a,1], Marie-Pierre Revel-Dubois [m,1]

[a] Radboud university medical center, P.O. Box 9101, 6500HB Nijmegen, The Netherlands
[b] University of Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany
[c] Independent researcher
[d] Keya medical technology co. ltd, Floor 20, Building A, 1 Ronghua South Road, Yizhuang Economic Development Zone, Daxing District, Beijing, PR China
[e] ML-Labs, Dublin City University, N210, Marconi building, Dublin City University, Glasnevin, Dublin 9, Ireland
[f] Faculty of Applied Computer Science, University of Augsburg, Germany
[g] Vrije Universiteit Brussel, Department of Electronics and Informatics, Pleinlaan 2, 1050 Brussels, Belgium
[h] Deakin University, Geelong, Australia
[i] University of New South Wales, Sydney, Australia
[j] IRCCS SYNLAB SDN, Naples, Italy
[k] Amazon Web Services, Marcel-Breuer-Str. 12, 80807 München, Germany
[l] TheraPanacea, 75004, Paris, France
[m] Department of Radiology, Université de Paris, APHP, Hôpital Cochin, 27 rue du Fg Saint Jacques, 75014 Paris, France
[n] imec, Kapeldreef 75, 3001 Leuven, Belgium

ARTICLE INFO

Keywords:
COVID-19
Machine learning
Medical image analysis challenge

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

* Corresponding author.
E-mail addresses: luuk.boulogne@radboudumc.nl (L.H. Boulogne), julian.lorenz@uni-a.de (J. Lorenz), simon.jegou.ia@gmail.com (S. Jégou), guangl@keyamedical.com (G. Li), mayug.maniparambil2@mail.dcu.ie (M. Maniparambil), miriam.elia@informatik.uni-augsburg.de (M. Elia), ine.dirks@vub.be (I. Dirks), imran.razzak@deakin.edu.au (I. Razzak), mario.verdicchio@synlab.it (M. Verdicchio).

[1] Challenge organizers.
[2] Team Code 1055.
[3] Team simon.j.
[4] Team Flying Bird.
[5] Team hal9000.
[6] Team uaux2.
[7] Team etro.
[8] Team deakin_team.
[9] Team SYNLAB-SDN.

https://doi.org/10.1016/j.media.2024.103230
Received 23 July 2023; Received in revised form 11 January 2024; Accepted 3 June 2024
Available online 5 June 2024

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.

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 solutions were readily available after the challenge had completed, Bulten et al. (2022), Aubreville et al. (2022), Schirmer et al. (2021), Da et al. (2022) and Ouyang et al. (2019) 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 (Prokop et al., 2020). Automatic algorithms that analyze CT scans of COVID-19 patients have the potential to aid healthcare professionals in the diagnostic process (Hassan et al., 2022). 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, Revel et al. (2021) a multicenter 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 2000 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. Fig. 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, Ouyang et al. (2019), Antonelli et al. (2021), Ehteshami Bejnordi et al. (2017), Lassau et al. (2020), Choi et al. (2022), Halabi et al. (2019), Ali et al. (2021), Knoll et al. (2020), Porwal et al. (2020), Kim et al. (2021), Fang et al. (2022), Sun et al. (2021), Sathianathen et al. (2022), Combalia et al. (2022), Kavur et al. (2021), Hakim et al. (2021), Heller et al. (2019), Bogunovic et al. (2019), Orlando et al. (2020), Yang et al. (2018), Hirvasniemi et al. (2023), Arganda-Carreras et al. (2015), Ivantsits et al. (2022), Caicedo et al. (2019), Simões et al. (2020), Veta et al. (2019), Winzeck et al. (2018), Marinescu et al. (2019), Balagurunathan et al. (2021), De Luca et al. (2021), Bratholm et al. (2021), Bron et al. (2015), Setio et al. (2017), Pan et al. (2019), Cash et al. (2015), Kim et al. (2020), Committee et al. (2021), Fu et al. (2020) and Babier et al. (2021) 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 (Bulten et al., 2022; Aubreville et al., 2022; Schirmer et al., 2021; Da et al., 2022; Sun et al., 2022; Hatt et al., 2018) solve this issue by requiring participants to submit functional algorithms. These can be made easily accessible to third parties (Bulten et al., 2022; Aubreville et al., 2022; Schirmer et al., 2021; Da et al., 2022), and generate reproducible results (Bulten et al., 2022; Sun et al., 2022).

We implement the Type Three (T3) challenge structure, which has only seen limited use in medical image analysis research (Schaffter et al., 2020). 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.

2. Materials and methods

2.1. Materials

Data from the STOIC study (Revel et al., 2021) 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 (Xie et al., 2020), 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. Fig. 2a depicts how the preprocessed database was split into training and evaluation sets for the Qualification and Final phases of STOIC2021.

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 (Reinke et al., 2021). 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.

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

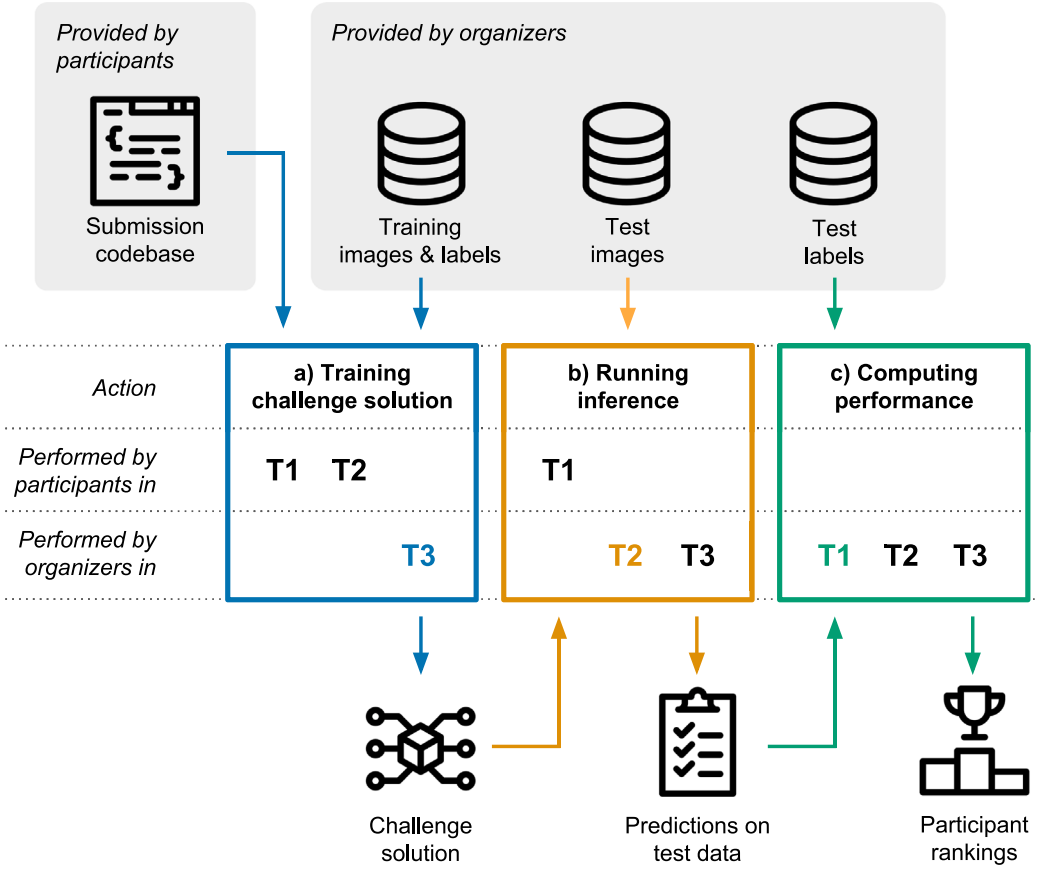


Fig. 1. Schematic representation of the submission pipeline of challenges of Type One (T1), Type Two (T2), and Type Three (T3). (a) A challenge solution is trained by applying a participants' codebase to images and labels provided by the challenge organizers. With T1 and T2, participants perform this 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.

in Fig. 2. These phases respectively followed the T2 and T3 format illustrated in Fig. 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.

2.3.1. 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 Fig. 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 Fig. 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.

2.3.2. 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 h with access to two Tesla V100 GPUs with 32 GB vRAM each, 16 cpus with a total of 128G RAM, and 2000 GB of Elastic Block Storage for storing intermediate results such as preprocessed data.

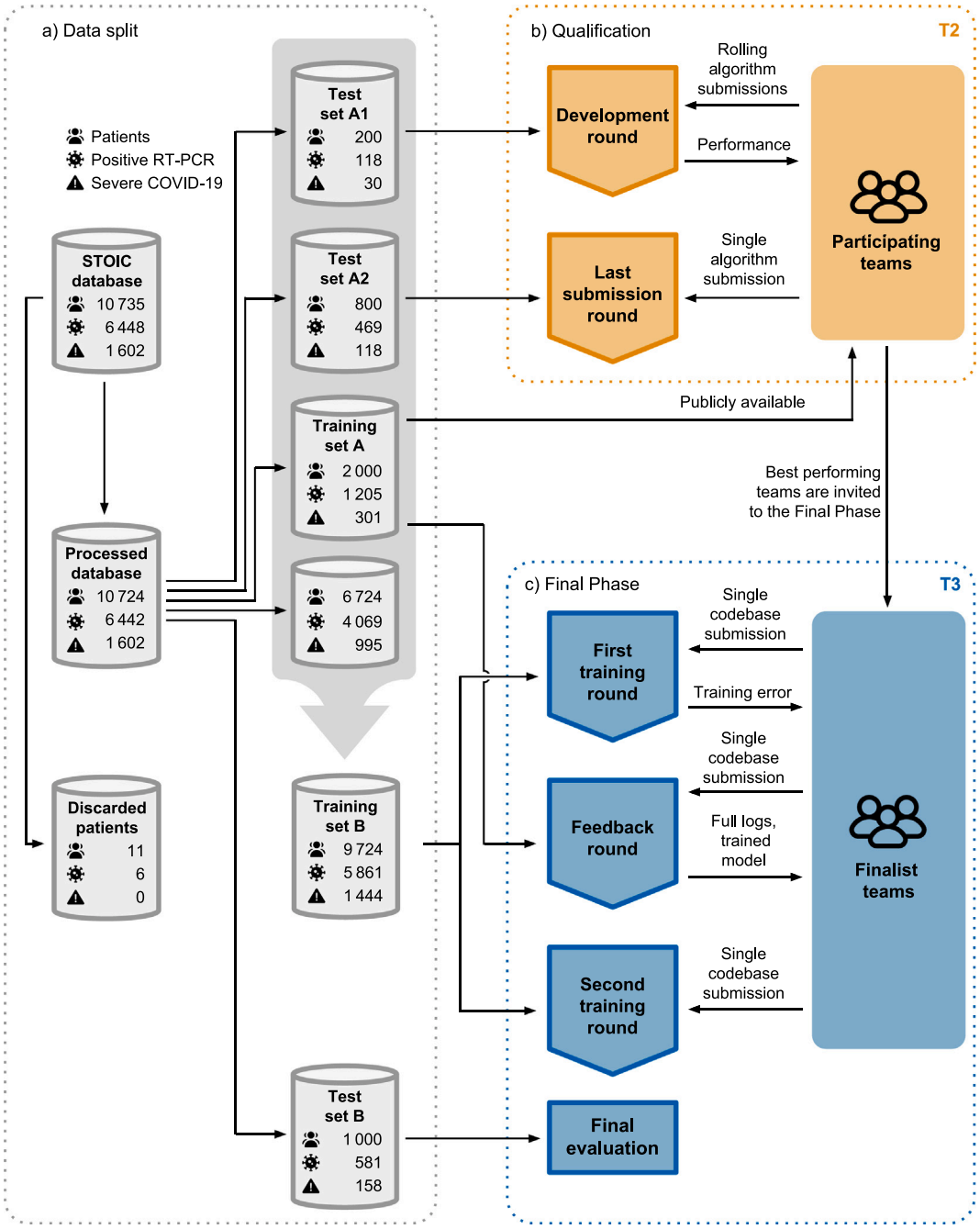


Fig. 2. Schematic overview of the STOIC2021 challenge. Each patient was represented by a single CT scan. (a) Schematic overview showing how many CT scans were used for what purpose, how many of them showed patients with a positive RT-PCR result, and how many of those patients suffered from severe COVID-19. 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.

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 (Merkel, 2014) container images from the submitted code bases and applied these to training set B (see Fig. 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

Table 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 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

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 h 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.

2.3.3. Prizes

Prizes in Amazon Web Services (AWS) credits were awarded to the best performing teams of the Final phase with values of $10 000, $6000, and $4000 for 1st, 2nd, and 3rd place respectively. The winners were announced during a public webinar on October 18th, 2022.

2.3.4. 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.

2.4. Statistical tests

The DeLong (DeLong et al., 1988; Sun and Xu, 2014) 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 1000 iterations (Moore and McCabe, 1989).

2.5. Baseline method

The baseline for STOIC2021 implemented a simple training and evaluation pipeline for an Inflated 3D convnet (I3D) (Carreira and Zisserman, 2017).

Preprocessing strategy. The input CT scans were resampled to an isotropic spacing of 1.6 mm^3. 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 (Carreira and Zisserman, 2017), 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 (Loshchilov and Hutter, 2018) 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.

3. Results

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. Fig. 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.

3.2. Final phase

3.2.1. 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 Fig. 3. In 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.

3.2.2. 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.

3.2.3. Performance

Table 1 shows the AUC on test set B for COVID-19 presence and severity of the teams that submitted to the Final phase. Fig. 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. Figs. 5 and 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. Figs. 7 and 8 highlight some individual cases from test set B. During the original STOIC project (Revel et al., 2021), 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).

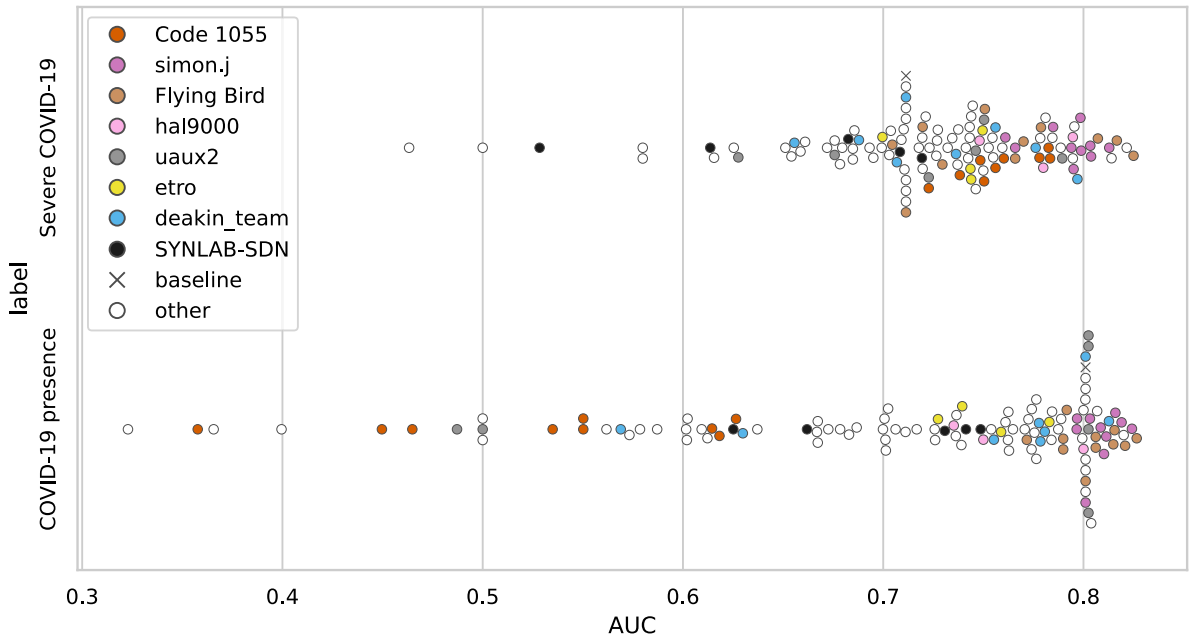


Fig. 3. Performance distribution of the rolling submissions to test set A1 during the Qualification phase. The performance of the baseline is represented by an 'x'. Submissions by the eight finalist teams are represented by colored circles. All other submissions are represented by white circles. Details about the metrics used are described in Section 2.2.

3.3. Solution methodology overview

Most finalists used lung and/or lesion segmentation methods (Hofmanninger et al., 2020; Müller et al., 2021) 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, Liu et al. (2022), He et al. (2016), Howard et al. (2019) and Huang et al. (2017) 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 (Dosovitskiy et al., 2020) (simon.j) or features designed based on generated lung (Hofmanninger et al., 2020) and lesion masks (NVIDIA NGC Catalog, 2023) (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.

3.3.1. Code 1055

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 (Liu et al., 2022) combines the benefits of the modern Vision Transformers (Dosovitskiy et al., 2020) 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 × 256 × 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 (Revel et al., 2021) is a comparably large database of CT scans, it is exceedingly small in contrast to ImageNet (Russakovsky et al., 2014). 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 (Roth et al., 2022; An et al., 2020; Clark et al., 2013) and classification (Morozov et al., 2020) dataset. We use an architecture inspired by UPerNet (Xiao et al., 2018) 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 Fig. 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. (2023).

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.

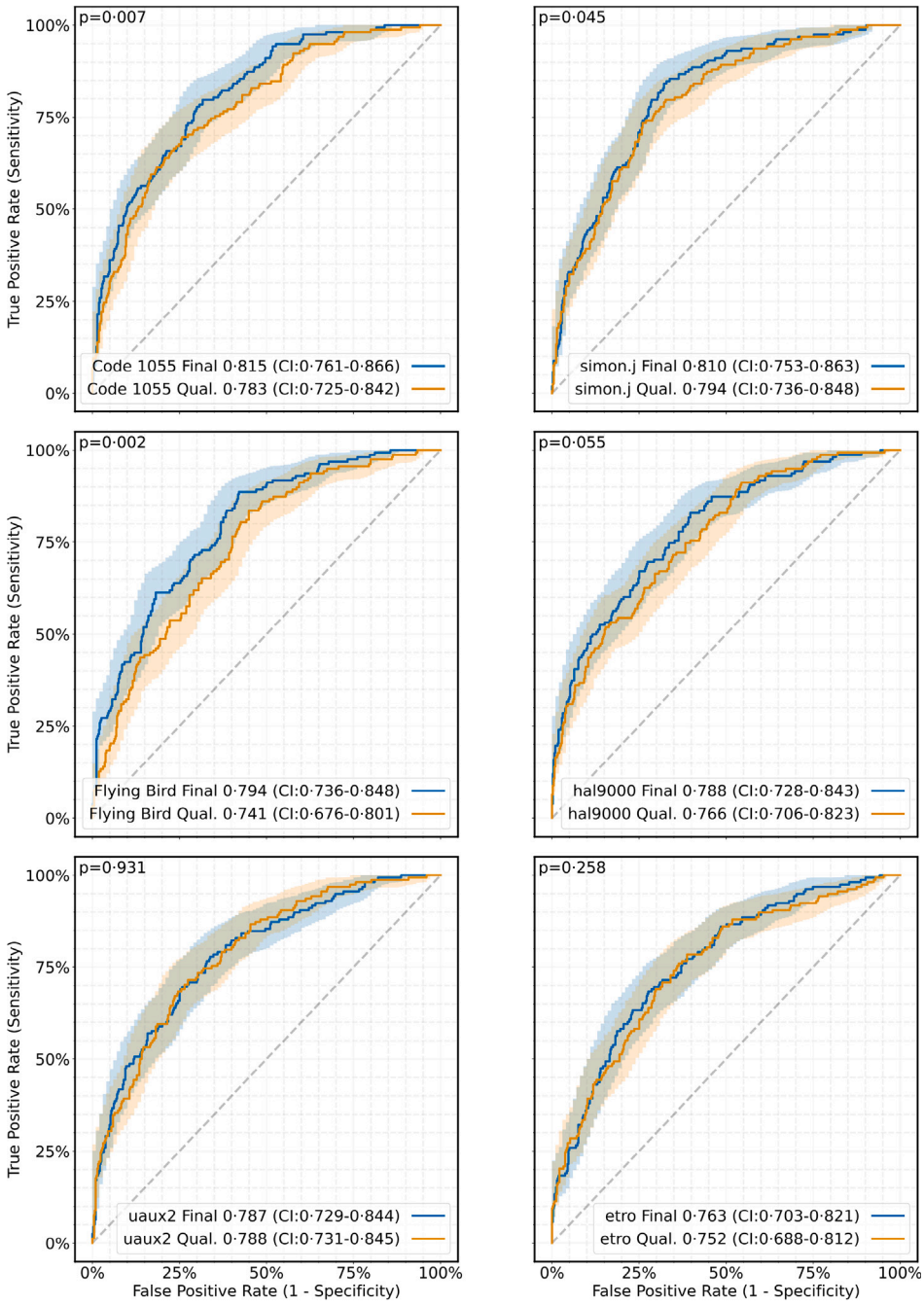


Fig. 4. ROC curves with confidence intervals (CIs) for discriminating between severe and non-severe COVID-19 on test set B. The curves for the codebase submissions in the Final phase that completed training on training set B successfully are shown in blue. The ROC curves of the submissions that represented 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.

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/.

3.3.2. simon.j

Balaitous is an updated version of the AI-severity algorithm (Lassau et al., 2021) implemented in the scancovia repository (Jégou, 2022). 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 × 1.5 mm × 5 mm and reshaped to a shape of 224 × 224 × 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 (Hofmanninger et al., 2020) 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 (Zhou et al., 2021). This model was pretrained on ImageNet-22k using iBOT (Zhou et al., 2021) 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

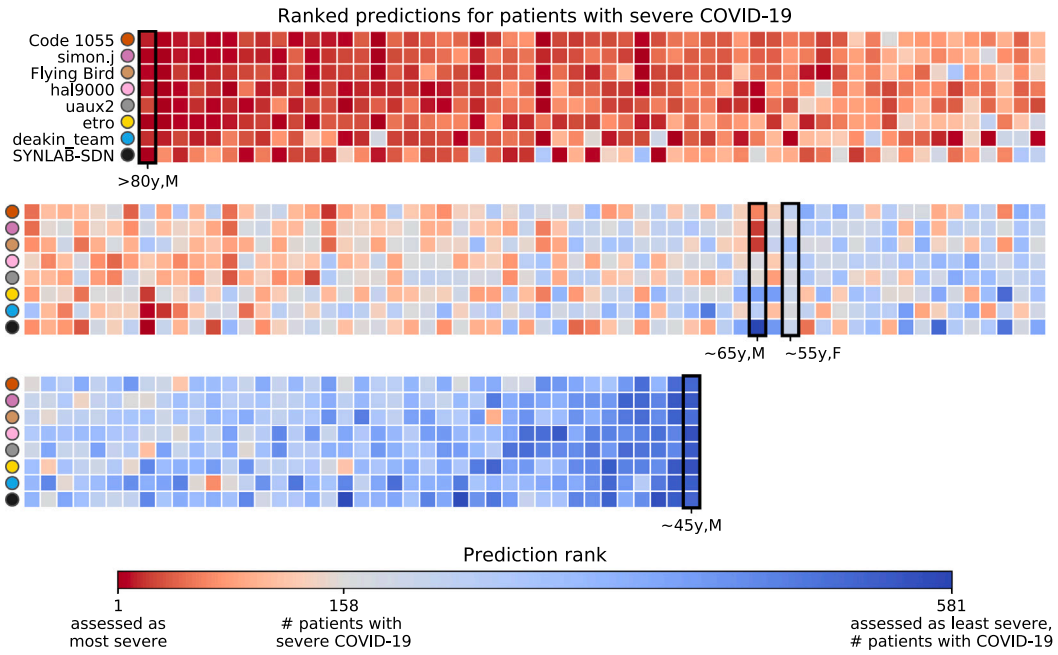


Fig. 5. Ranked predictions for the subjects with severe COVID-19. Ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19. Each column shows the ranked predictions of all finalist teams for one subject. The subjects are ordered by the average rank of all corresponding finalist predictions. Fig. 7 shows the CT scans corresponding to the columns that are outlined in black and annotated with age and sex.

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/.

3.3.3. Flying bird

The method employed was end-to-end deep learning with ResNet18 (He et al., 2016) models.

Preprocessing strategy. In order to minimize image size and eliminate irrelevant regions, an open source lung segmentation model (Hofmanninger et al., 2020) 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 (Zhou et al., 2019), 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 He et al. (2016), 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 (Loshchilov and Hutter, 2018) 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/.

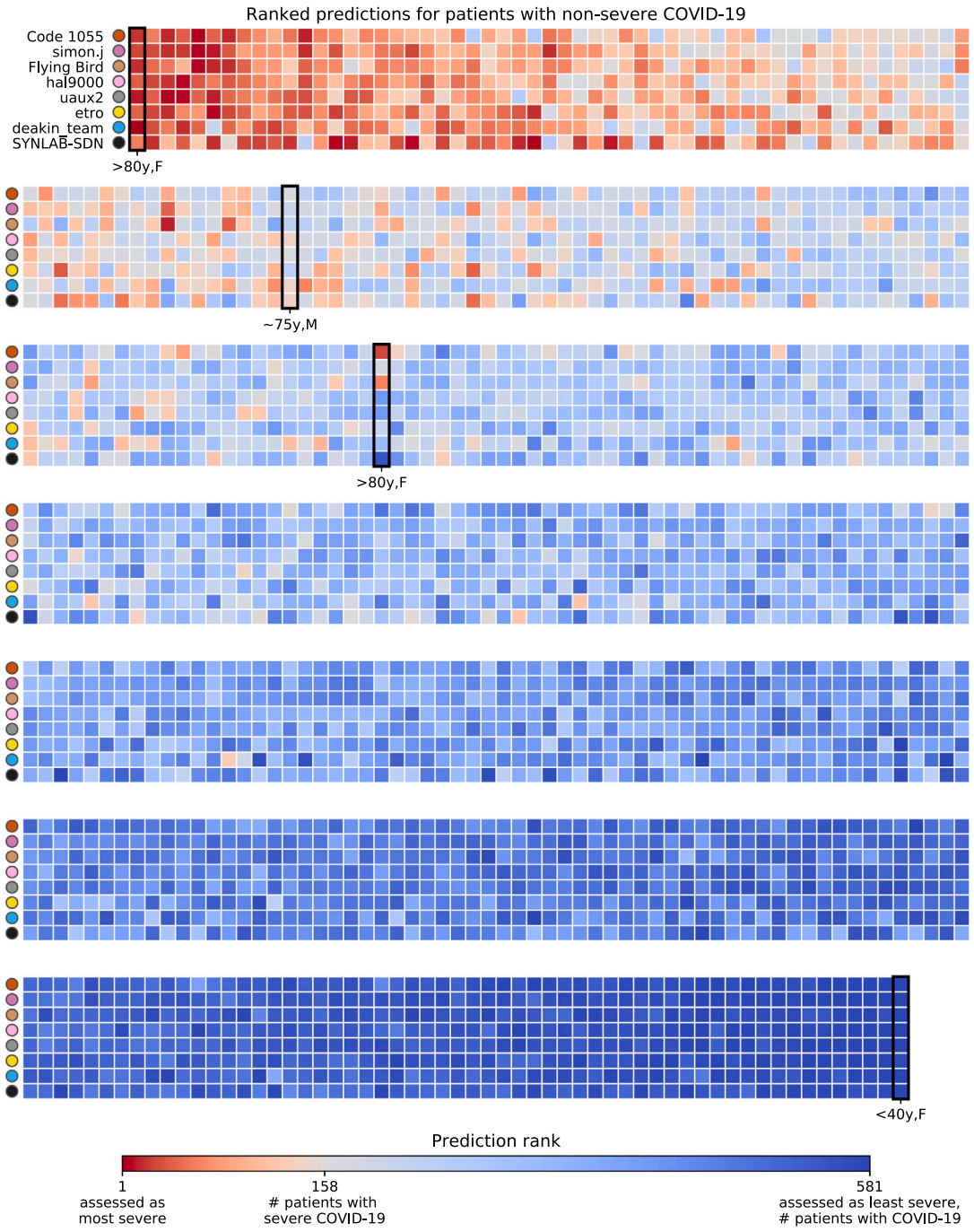


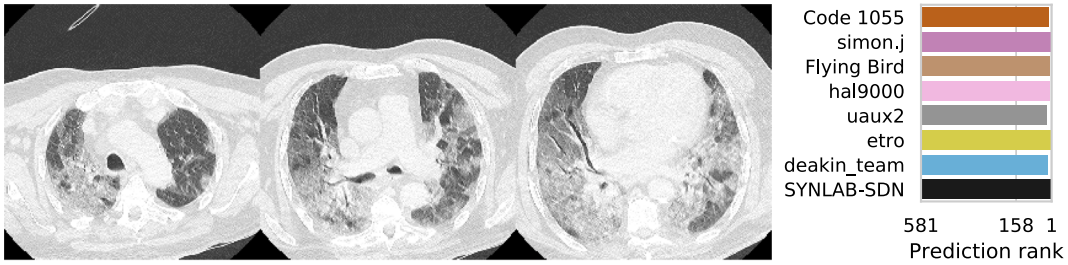
Fig. 6. Ranked predictions for severe COVID-19 for subjects with non-severe COVID-19. Ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19. Each column shows the ranked predictions of all finalist teams for one subject. The subjects are ordered by the average rank of all corresponding finalist predictions. Fig. 8 shows the CT scans corresponding to the columns that are outlined in black and annotated with age and sex.

3.3.4. hal9000

We employed an ensemble of ResNet18 (He et al., 2016), and MoblieNetV3-Large (Howard et al., 2019) 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.

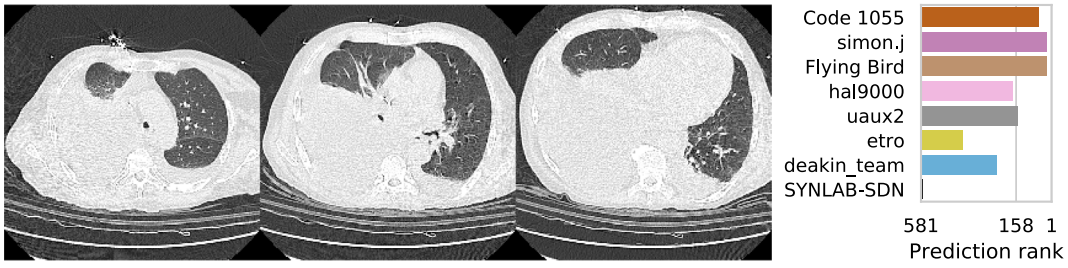
Preprocessing strategy. 32 equidistant slices were sampled from the input CT scan. These slices were resampled to 224 × 224 pixels. The pixel values were clipped between −1350 and 150 HU. The images were normalized to a mean of 0.5 and a standard deviation of 0.5.

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 (He et al., 2016) was trained on five of these splits, and a MobileNetV3-Large (Howard et al., 2019) 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.

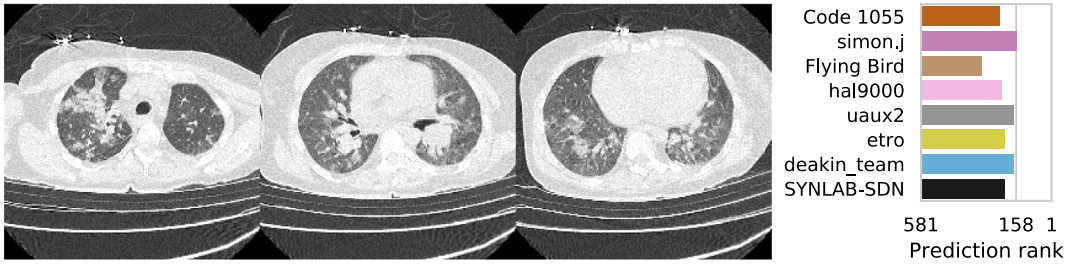


(a) 80+ years old, male

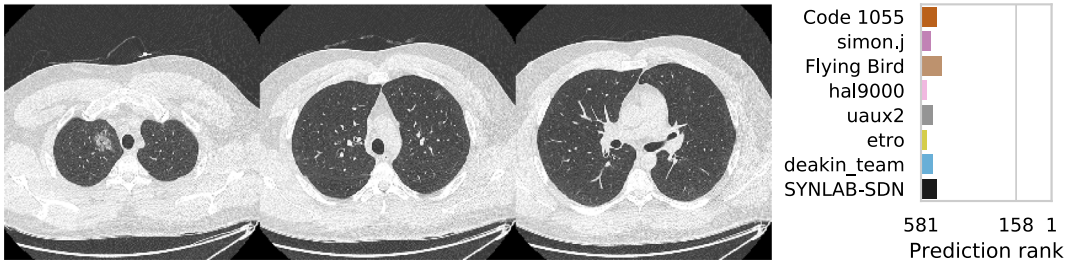
article



(b) 60-70 years old, male



(c) 50-60 years old, female



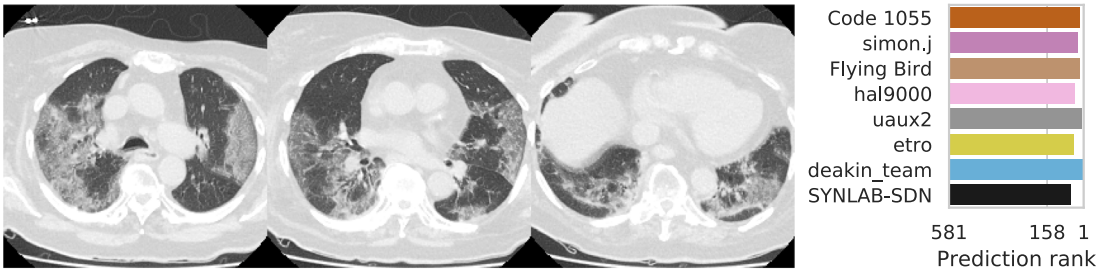
(d) 40-50 years old, male

Fig. 7. Subjects from test set B with severe COVID-19 that were highlighted in Fig. 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.

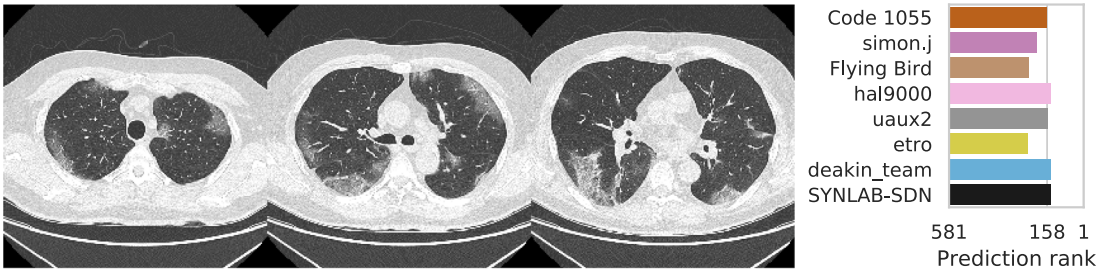
Inference strategy. All model predictions were combined through averaging. We employed extensive test time augmentations involving five different crops (four corner crops and the center crop), and three different rotations (minus five degrees, plus five and plus ten degrees), and averaged the predictions 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.

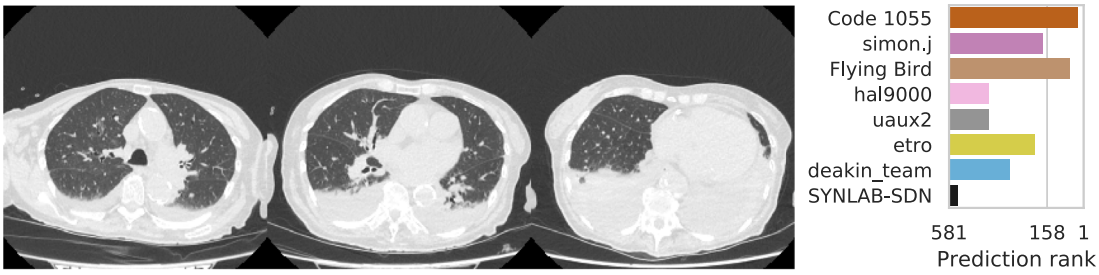
Public access. Code for training and inference publicly available at https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-hal9000. Algorithm available for public use at https://grand-challenge.org/algorithms/hal9000-second-final-phase-submission/.



(a) 80+ years old, female



(b) 70-80 years old, male



(c) 80+ years old, female

(d) 40- years old, female

Fig. 8. Subjects from test set B with non-severe COVID-19 that were highlighted in Fig. 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.

3.3.5. 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 (Müller and Kramer, 2022).

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 (Yamada et al., 2022). 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 of [0, 1]. Then, normalization

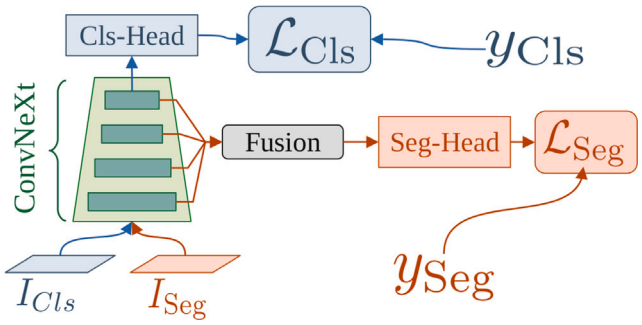


Fig. 9. The pretraining pipeline is depicted. If segmentation data (I_{Seg}) is used as input, the features of each stage are upsampled, concatenated and the segmentation map is calculated with a segmentation head. If the classification data (I_{Cls}) is used as input, the severity prediction is obtained with a classification head using the features of the last stage. The overall loss is calculated as $L = L_{Cls} + L_{Seg}$.

was applied via the Z-Score normalization approach based on the mean and standard deviation computed on the ImageNet dataset (Deng et al., 2009).

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 (Lin et al., 2017). 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 presence 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 (Müller et al., 2021; Müller and Kramer, 2021), which utilizes a standard U-Net to predict infected areas Fig. 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/.

3.3.6. 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 (Hofmanninger et al., 2020). 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 (MON, 2022) was used. The hyperparameters for this deep learning pipeline were determined automatically using the heuristics developed in nnU-Net (Isensee et al., 2019). 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 (COV, 2022), 69 scans and manual lung lesion segmentations from the icovid consortium (ICo, 2022), 70 scans from the COPLENet public dataset (COP, 2022) and 10 scans from the publicly available COVID-19 CT Lung and Infection Segmentation Dataset (Ma et al., 2020). 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 (Liu et al., 2022) 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/.

3.3.7. deakin_team

The method employed was end-to-end deep learning with DenseNet-201 (Huang et al., 2017).

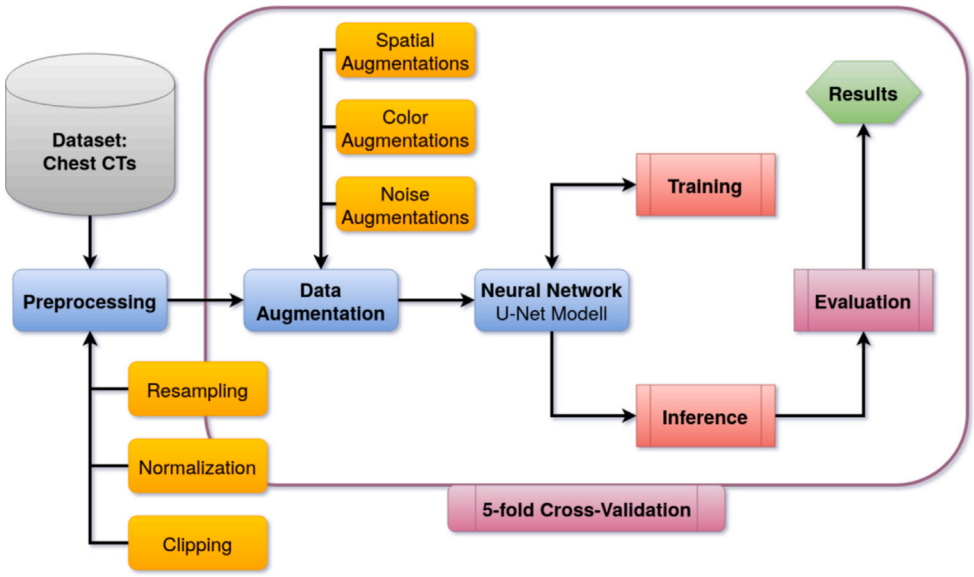


Fig. 10. The MIScnn pipeline for SARS-CoV-2 segmentation to calculate the Infection-Lung-Ratio (Müller et al., 2021).

Preprocessing strategy. The input CT scans were resampled to an isotropic spacing of 1.6 mm^3. 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 (Huang et al., 2017), 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, DenseNet-169, and DenseNet-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).

3.3.8. 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 (Hofmanninger et al., 2020). 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. 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 Ouyang et al. (2019), Antonelli et al. (2021), Ehteshami Bejnordi et al. (2017), Lassau et al. (2020), Choi et al. (2022), Halabi et al. (2019), Ali et al. (2021), Knoll et al. (2020), Porwal et al. (2020), Kim et al. (2021), Fang et al. (2022), Sun et al. (2021), Sathianathen et al. (2022), Combalia et al. (2022), Kavur et al. (2021), Hakim et al. (2021), Heller et al. (2019), Bogunovic et al. (2019), Orlando et al. (2020), Yang et al. (2018), Hirvasniemi et al. (2023), Arganda-Carreras et al. (2015), Ivantsits et al. (2022), Caicedo et al. (2019), Simões et al. (2020), Veta et al. (2019), Winzeck et al. (2018), Marinescu et al. (2019), Balagurunathan et al. (2021), De Luca et al. (2021), Bratholm et al. (2021), Bron et al. (2015), Setio et al. (2017), Pan et al. (2019), Cash et al. (2015), Kim et al. (2020), Committee et al. (2021), Fu et al. (2020) and Babier et al. (2021). STOIC2021 consisted of a Qualification phase that instead followed the structure implemented of some recent challenges (Bulten et al., 2022; Aubreville et al., 2022; Schirmer et al., 2021; Da et al., 2022; Sun et al., 2022; Hatt et al., 2018) 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 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 (Revel et al., 2021)

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 versatility 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 (Tsang et al., 2021). 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 outcome by Lassau et al. (2021). 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 2.3.2, 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 (Revel et al., 2021).

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.

CRediT authorship contribution statement

Luuk H. Boulogne: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Julian Lorenz:** Investigation, Methodology, Software, Writing – review & editing. **Daniel Kienzle:** Investigation, Methodology, Software, Writing – review & editing. **Robin Schön:** Investigation, Methodology, Software, Writing – review & editing. **Katja Ludwig:** Investigation, Methodology, Software, Writing – review & editing. **Rainer Lienhart:** Investigation, Methodology, Software, Writing – review & editing. **Simon Jégou:** Investigation, Methodology, Software, Writing – review & editing. **Guang Li:** Investigation, Methodology, Software, Writing – review & editing. **Cong Chen:** Investigation, Methodology, Software, Writing – review & editing. **Qi Wang:** Investigation, Methodology, Software, Writing – review & editing. **Mayug Maniparambil:** Investigation, Methodology, Software, Writing – review & editing. **Dominik Müller:** Investigation, Methodology, Software, Writing – review & editing. **Silvan Mertes:** Investigation, Methodology, Software, Writing – review & editing. **Niklas Schröter:** Investigation, Methodology, Software, Writing – review & editing. **Fabio Hellmann:** Investigation, Methodology, Software, Writing – review & editing. **Miriam Elia:** Investigation, Methodology, Software, Writing – review & editing. **Ine Dirks:** Investigation, Methodology, Software, Writing – review & editing. **Matías Nicolás Bossa:** Investigation, Methodology, Software, Writing – review & editing. **Abel Díaz Berenguer:** Investigation, Methodology, Software, Writing – review & editing. **Tanmoy Mukherjee:** Investigation, Methodology, Software, Writing – review & editing. **Jef Vandemeulebroucke:** Investigation, Methodology, Software, Writing – review & editing. **Hichem Sahli:** Investigation, Methodology, Software, Writing – review & editing. **Nikos Deligiannis:** Investigation, Methodology, Software, Writing – review & editing. **Panagiotis Gonidakis:** Investigation, Methodology, Software, Writing – review & editing. **Ngoc Dung Huynh:** Investigation, Methodology, Software, Writing – review & editing. **Imran Razzak:** Investigation, Methodology, Software, Writing – review & editing. **Reda Bouadjenek:** Conceptualization, Investigation, Methodology, Software, Writing – review & editing. **Mario Verdicchio:** Investigation, Methodology, Software, Writing – review & editing. **Pasquale Borrelli:** Investigation, Methodology, Software, Writing – review & editing. **Marco Aiello:** Investigation, Methodology, Software, Writing – review & editing. **James A. Meakin:** Investigation, Project administration, Resources, Software, Validation, Writing – review & editing. **Alexander Lemm:** Funding acquisition, Writing – review & editing. **Christoph Russ:** Funding acquisition, Writing – review & editing. **Razvan Ionasec:** Funding acquisition, Writing – review & editing. **Nikos Paragios:** Writing – review & editing. **Bram van Ginneken:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Marie-Pierre Revel-Dubois:** Data curation, Supervision, Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

2000 CT scans were publicly released. See the data statement for access to the full STOIC database.

Declaration of generative AI in scientific writing

During the preparation of this work the author(s) used ChatGPT in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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 (Revel et al., 2021) 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.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103230.

References

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. 70, 102002.

An, P., Xu, S., Harmon, S., Turkbey, E., Sanford, T., Amalou, A., et al., 2020. Ct images in covid-19. The cancer imaging archive.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., AnnetteKopp-Schneider, Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Huisman, H., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbelaez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, N., Kim, I., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2021. The medical segmentation decathlon. arXiv preprint arXiv: 2106.05735. URL: https://arxiv.org/abs/2106.05735.

Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Cireşan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J.M., et al., 2015. Crowdsourcing the creation of image segmentation algorithms for connectomics. Front. Neuroanat. 9, 142.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., Breen, J., Ravikumar, N., Chung, Y., Park, J., Nateghi, R., Pourakpour, F., Fick, R.H.J., Ben Hadj, S., Jahanifar, M., Shephard, A., Dexl, J., Wittenberg, T., Kondo, S., Lafarge, M.W., Koelzer, V.H., Liang, J., Wang, Y., Long, X., Liu, J., Razavi, S., Khademi, A., Yang, S., Wang, X., Erber, R., Klang, A., Lipnik, K., Bolfa, P., Dark, M.J., Wasinger, G., Veta, M., Breininger, K., 2022. Mitosis domain generalization in histopathology images - The MIDOG challenge. Med. Image Anal. (ISSN: 1361-8423) 84, 102699. http://dx.doi.org/10.1016/j.media.2022.102699.

Babier, A., Zhang, B., Mahmood, R., Moore, K.L., Purdie, T.G., McNiven, A.L., Chan, T.C., 2021. OpenKBP: the open-access knowledge-based planning grand challenge and dataset. Med. Phys. 48 (9), 5549–5561.

Balagurunathan, Y., Beers, A., Mcnitt-Gray, M., Hadjiiski, L., Napel, S., Goldgof, D., Perez, G., Arbelaez, P., Mehrtash, A., Kapur, T., et al., 2021. Lung nodule malignancy prediction in sequential ct scans: Summary of isbi 2018 challenge. IEEE Trans. Med. Imaging 40 (12), 3748–3761.

Bogunovic, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., Gopinath, K., Gostar, A.K., Jeon, K., Ji, Z., Kang, S.H., Koozekanani, D.D., Lu, D., Morley, D., Parhi, K.K., Park, H.S., Rashno, A., Sarunic, M., Shaikh, S., Sivaswamy, J., Tennakoon, R., Yadav, S., De Zanet, S., Waldstein, S.M., Gerendas, B.S., Klaver, C., Sánchez, C.I., Schmidt-Erfurth, U., 2019. RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imaging 38, 1858–1874. http://dx.doi.org/10.1109/TMI.2019.2901398.

Bratholm, L.A., Gerrard, W., Anderson, B., Bai, S., Choi, S., Dang, L., Hanchar, P., Howard, A., Kim, S., Kolter, Z., et al., 2021. A community-powered search of machine learning strategy space to find NMR property prediction models. PLoS ONE 16 (7), e0253612.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the caddementia challenge. NeuroImage 111, 562–579.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Strom, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Gronberg, H., Samaratunga, H., Delahunt, B., Tsuzuki, T., Hakkinen, T., Egevad, L., Demkin, M., Dane, S., Tan, F., Valkonen, M., Corrado, G.S., Peng, L., Mermel, C.H., Ruusuvuori, P., Litjens, G., Eklund, M., Brilhante, A., Cakir, A., Farre, X., Geronatsiou, K., Molinie, V., Pereira, G., Roy, P., Saile, G., Salles, P.G.O., Schaafsma, E., Tschui, J., Billoch-Lima, J., Pereira, E.M., Zhou, M., He, S., Song, S., Sun, Q., Yoshihara, H., Yamaguchi, T., Ono, K., Shen, T., Ji, J., Roussel, A., Zhou, K., Chai, T., Weng, N., Grechka, D., Shugaev, M.V., Kiminya, R., Kovalev, V., Voynov, D., Malyshev, V., Lapo, E., Campos, M., Ota, N., Yamaoka, S., Fujimoto, Y., Yoshioka, K., Juvonen, J., Tukiainen, M., Karlsson, A., Guo, R., Hsieh, C.-L., Zubarev, I., Bukhar, H.S.T., Li, W., Li, J., Speier, W., Arnold, C., Kim, K., Bae, B., Kim, Y.W., Lee, H.-S., Park, J., the PANDA challenge consortium, 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. http://dx.doi.org/10.1038/s41591-021-01620-2.

Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al., 2019. Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat. Methods 16 (12), 1247–1253.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, http://dx.doi.org/10.1109/cvpr.2017.502.

Cash, D.M., Frost, C., Iheme, L.O., Ünay, D., Kandemir, M., Fripp, J., Salvado, O., Bourgeat, P., Reuter, M., Fischl, B., et al., 2015. Assessing atrophy measurement techniques in dementia: Results from the MIRIAD atrophy challenge. Neuroimage 123, 149–164.

Choi, H., Kim, H., Jin, K.N., Jeong, Y.J., Chae, K.J., Lee, K.H., Yong, H.S., Gil, B., Lee, H.-J., Lee, K.Y., et al., 2022. A challenge for emphysema quantification using a deep learning algorithm with low-dose chest computed tomography. J. Thorac. Imaging 37 (4), 253–261.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J. Digit. Imaging 26 (6), 1045–1057. http://dx.doi.org/10.1007/s10278-013-9622-7.

Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N.R., Liopyris, K., et al., 2022. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. Lancet Digit. Health 4 (5), e330–e339.

Committee, Q.C..O., Bilgic, B., Langkammer, C., Marques, J.P., Meineke, J., Milovic, C., Schweser, F., 2021. QSM reconstruction challenge 2.0: Design and report of results. Magn. Reson. Med. 86 (3), 1241–1255.

2022. COPLE-net. URL: https://github.com/HiLab-git/COPLE-Net. (Accessed 12 May 2022).

2022. COVID-19 lung CT lesion segmentation grand challenge. URL: https://covid-segmentation.grand-challenge.org/. (Accessed 12 May 2022).

Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., et al., 2022. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Med. Image Anal. 80, 102485.

De Luca, A., Ianus, A., Leemans, A., Palombo, M., Shemesh, N., Zhang, H., Alexander, D.C., Nilsson, M., Froeling, M., Biessels, G.-J., et al., 2021. On the generalizability of diffusion MRI signal representations across acquisition parameters, sequences and tissue types: Chronicles of the MEMENTO challenge. NeuroImage 240, 118367.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ehteshami Bejnordi, B., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., the CAMELYON16 Consortium, Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M.Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. J. Am. Med. Assoc. 318, 2199–2210. http://dx.doi.org/10.1001/jama.2017.14585.

Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., et al., 2022. ADAM challenge: detecting age-related macular degeneration from fundus images. IEEE Trans. Med. Imaging 41 (10), 2828–2847.

Fu, H., Li, F., Sun, X., Cao, X., Liao, J., Orlando, J.I., Tao, X., Li, Y., Zhang, S., Tan, M., et al., 2020. Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography. Med. Image Anal. 66, 101798.

Hakim, A., Christensen, S., Winzeck, S., Lansberg, M.G., Parsons, M.W., Lucas, C., Robben, D., Wiest, R., Reyes, M., Zaharchuk, G., 2021. Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the ISLES challenge. Stroke 52 (7), 2328–2337.

Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., et al., 2019. The RSNA pediatric bone age machine learning challenge. Radiology 290 (2), 498–503.

Hassan, H., Ren, Z., Zhao, H., Huang, S., Li, D., Xiang, S., Kang, Y., Chen, S., Huang, B., 2022. Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. Comput. Biol. Med. 141, 105123.

Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S., Li, L., Lu, W., Jaouen, V., Tauber, C., Czakon, J., et al., 2018. The first MICCAI challenge on PET tumor segmentation. Med. Image Anal. 44, 177–195.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 770–778. http://dx.doi.org/10.1109/cvpr.2016.90.

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

Hirvasniemi, J., Runhaar, J., van der Heijden, R., Zokaeinikoo, M., Yang, M., Li, X., Tan, J., Rajamohan, H., Zhou, Y., Deniz, C., et al., 2023. The knee OsteoArthritis prediction (KNOAP2020) challenge: An image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. Osteoarthr. Cartil. 31 (1), 115–125.

Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G., 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. 4 (1), 1–13.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

2022. iCOVID AI. URL: https://icovid.ai/. (Accessed 12 May 2022).

Isensee, F., Petersen, J., Kohl, S.A.A., Jäger, P.F., Maier-Hein, K.H., 2019. Nnu-net: Breaking the spell on successful medical image segmentation. arXiv:1904.08128. arXiv:http://arxiv.org/abs/1904.08128v1.

Ivantsits, M., Goubergrits, L., Kuhnigk, J.-M., Huellebrand, M., Bruening, J., Kossen, T., Pfahringer, B., Schaller, J., Spuler, A., Kuehne, T., et al., 2022. Detection and analysis of cerebral aneurysms based on X-ray rotational angiography-the CADA 2020 challenge. Med. Image Anal. 77, 102333.

Jégou, S., 2022. Scancovia repository. URL: https://github.com/owkin/scancovia/tree/main/. (Accessed 20 December 2022).

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950.

Kienzle, D., Lorenz, J., Schön, R., Ludwig, K., Lienhart, R., 2023. 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, pp. 500–516.

Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al., 2021. PAIP 2019: Liver cancer segmentation challenge. Med. Image Anal. 67, 101854.

Kim, Y.-G., Song, I.H., Lee, H., Kim, S., Yang, D.H., Kim, N., Shin, D., Yoo, Y., Lee, K., Kim, D., et al., 2020. 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. Cancer Res. Treat.: Off. J. Korean Cancer Assoc. 52 (4), 1103–1111.

Knoll, F., Murrell, T., Sriram, A., Yakubova, N., Zbontar, J., Rabbat, M., Defazio, A., Muckley, M.J., Sodickson, D.K., Zitnick, C.L., et al., 2020. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. Magn. Reson. Med. 84 (6), 3054–3070.

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., Soliman, S., Meyrignac, O., Talabard, M.-P., Lamarque, J.-P., et al., 2021. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat. Commun. 12 (1), 1–11.

Lassau, N., Bousaid, I., Chouzenoux, E., Lamarque, J.-P., Charmettant, B., Azoulay, M., Cotton, F., Khalil, A., Lucidarme, O., Pigneur, F., et al., 2020. Three artificial intelligence data challenges based on CT and MRI. Diagn. Interv. Imaging 101 (12), 783–788.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. arXiv:http://arxiv.org/abs/1708.02002v2.

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

Loshchilov, I., Hutter, F., 2018. Fixing weight decay regularization in adam. In: Proceedings of the ICLR 2018 Conference Blind. URL: http://arxiv.org/abs/1711.05101. CoRR, abs/1711.05101.

Ma, J., Ge, C., Wang, Y., An, X., Gao, J., Yu, Z., Zhang, M., Liu, X., Deng, X., Cao, S., Wei, H., Mei, S., Yang, X., Nie, Z., Li, C., Tian, L., Zhu, Y., Zhu, Q., Dong, G., He, J., 2020. COVID-19 CT lung and infection segmentation dataset. http://dx.doi.org/10.5281/zenodo.3757476.

Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Golland, P., Klein, S., et al., 2019. 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, pp. 1–10.

Merkel, D., 2014. Docker: Lightweight linux containers for consistent development and deployment. https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf. (Accessed 5 December 2022).

2022. MONAI documentation - DynUNet. URL: https://docs.monai.io/en/stable/networks.html#dynunet. (Accessed 12 May 2022).

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

Morozov, S.P., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I.A., Gelezhe, P., Gonchar, A., Chernina, V.Y., 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465.

Müller, D., Kramer, F., 2021. MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning. BMC Med. Imaging 21 (1), 1–11.

Müller, D., Kramer, F., 2022. AUCMEDI - a framework for automated classification of medical images. URL: https://github.com/frankkramer-lab/aucmedi. (Accessed 12 May 2022).

Müller, D., Soto-Rey, I., Kramer, F., 2021. Robust chest CT image segmentation of COVID-19 lung infection based on limited data. Inform. Med. Unlocked 25, 100681.

NVIDIA NGC Catalog, 2023. Clara_train_covid19_ct_lesion_seg. https://catalog.ngc.nvidia.com/orgs/nvidia/models/clara_train_covid19_ct_lesion_seg.

Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.-A., Kim, J., Lee, J., et al., 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. 59, 101570.

Ouyang, W., Winsnes, C.F., Hjelmare, M., Cesnik, A.J., Åkesson, L., Xu, H., Sullivan, D.P., Dai, S., Lan, J., Jinmo, P., et al., 2019. Analysis of the human protein atlas image classification competition. Nat. Methods 16 (12), 1254–1261.

Pan, I., Thodberg, H.H., Halabi, S.S., Kalpathy-Cramer, J., Larson, D.B., 2019. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. Radiol.: Artif. Intell. 1 (6), e190053.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al., 2020. Idrid: Diabetic retinopathy–segmentation and grading challenge. Med. Image Anal. 59, 101561.

Prokop, M., van Everdingen, W., van Rees Vellinga, T., Quarles van Ufford, J., Stoger, L., Beenen, L., Geurts, B., Gietema, H., Krdzalic, J., Schaefer-Prokop, C., van Ginneken, B., Brink, M., the COVID-19 Standardized Reporting Working Group of the Dutch Radiological Society, 2020. CO-RADS - a categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation. Radiology 296 (2), E97–E104. http://dx.doi.org/10.1148/radiol.2020201473.

Reinke, A., Tizabi, M.D., Sudre, C.H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., et al., 2021. Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642.

Revel, M.-P., Boussouar, S., de Margerie-Mellon, C., Saab, I., Lapotre, T., Mompoint, D., Chassagnon, G., Milon, A., Lederlin, M., Bennani, S., et al., 2021. Study of thoracic CT in COVID-19: the STOIC project. Radiology 301 (1), E361–E370.

Roth, H.R., Xu, Z., Tor-Díez, C., Jacob, R.S., Zember, J., Molto, J., Li, W., Xu, S., Turkbey, B., Turkbey, E., et al., 2022. Rapid artificial intelligence solutions in a pandemic—The COVID-19-20 lung CT lesion segmentation challenge. Med. Image Anal. 82, 102605.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2014. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 1–42.

Sathianathen, N.J., Heller, N., Tejpaul, R., Stai, B., Kalapara, A., Rickman, J., Dean, J., Oestreich, M., Blake, P., Kaluzniak, H., et al., 2022. Automatic segmentation of kidneys and kidney tumors: The KiTS19 international challenge. Front. Digit. Health 3, 797607.

Schaffter, T., Buist, D.S., Lee, C.I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S., et al., 2020. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw. Open 3 (3), e200265.

Schirmer, M.D., Venkataraman, A., Rekik, I., Kim, M., Mostofsky, S.H., Nebel, M.B., Rosch, K., Seymour, K., Crocetti, D., Irzan, H., et al., 2021. Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge. Med. Image Anal. 70, 101972.

Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., Bogaard, C.v.d., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., Gugten, R.v.d., Heng, P.A., Jansen, B., de Kaste, M.M.J., Kotov, V., Lin, J.Y.-H., Manders, J.T.M.C., Sonora-Mengana, A., Garcia-Naranjo, J.C., Papavasileiou, E., Prokop, M., Saletta, M., Schaefer-Prokop, C.M., Scholten, E.T., Scholten, L., Snoeren, M.M., Torres, E.L., Vandemeulebroucke, J., Walasek, N., Zuidhof, G.C.A., Ginneken, B.v., Jacobs, C., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med. Image Anal. 42, 1–13. http://dx.doi.org/10.1016/j.media.2017.06.015, URL: https://arxiv.org/abs/1612.08012.

Simões, M., Borra, D., Santamaría-Vázquez, E., GBT-UPM, Bittencourt-Villalpando, M., Krzemiński, D., Miladinović, A., Neural Engineering Group, Schmid, T., Zhao, H., et al., 2020. BCIAUT-P300: A multi-session and multi-subject benchmark dataset on autism for P300-based brain-computer-interfaces. Front. Neurosci. 14, 568104.

Sun, Y., Gao, K., Wu, Z., Li, G., Zong, X., Lei, Z., Wei, Y., Ma, J., Yang, X., Feng, X., et al., 2021. Multi-site infant brain segmentation algorithms: the iSeg-2019 challenge. IEEE Trans. Med. Imaging 40 (5), 1363–1376.

Sun, D., Nguyen, T.M., Allaway, R.J., Wang, J., Chung, V., Thomas, V.Y., Mason, M., Dimitrovsky, I., Ericson, L., Li, H., et al., 2022. A crowdsourcing approach to develop machine learning models to quantify radiographic joint damage in rheumatoid arthritis. JAMA Netw. Open 5 (8), e2227423.

Sun, X., Xu, W., 2014. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process. Lett. 21 (11), 1389–1393.

Tsang, N.N.Y., So, H.C., Ng, K.Y., Cowling, B.J., Leung, G.M., Ip, D.K.M., 2021. Diagnostic performance of different sampling approaches for SARS-CoV-2 RT-PCR testing: a systematic review and meta-analysis. Lancet Infect. Dis. 21 (9), 1233–1245.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjoblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.-C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P.W., 2019. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med. Image Anal. 54 (5), 111–121. http://dx.doi.org/10.1016/j.media.2019.02.012.

Winzeck, S., Hakim, A., McKinley, R., Pinto, J.A., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., et al., 2018. ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Front. Neurol. 9, 679.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 418–434.

Xie, W., Jacobs, C., Charbonnier, J.-P., van Ginneken, B., 2020. Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans. IEEE Trans. Med. Imaging 39, 2664–2675. http://dx.doi.org/10.1109/TMI.2020.2995108.

Yamada, D., Ohde, S., Imai, R., Ikejima, K., Matsusako, M., Kurihara, Y., 2022. Visual classification of three computed tomography lung patterns to predict prognosis of COVID-19: a retrospective study. BMC Pulm. Med. 22, 1–9.

Yang, J., Veeraraghavan, H., Armato, III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., et al., 2018. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med. Phys. 45 (10), 4568–4581.

Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. 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, pp. 384–393.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2021. Ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832.