Medical image interpretation challenges and research activities of the tAImedIA group at UniBS

Alberto Signoroni[1,*], Mattia Savardi[1,*], Davide Farina[1], Sergio Benini[2], Edoardo Coppola[2], Damiano Ferrari[2], Mauro Massussi[3], Salvatore Curello[3], Michele Svanera[4] and Giuseppe D'Ancona[5]

[1]University of Brescia, Department of Medical and Surgical Specialities, Radiological Sciences and Public Health, Brescia, Italy
[2]University of Brescia, Department of Information Engineering, Brescia, Italy
[3]ASST Spedali Civili di Brescia, Department of Cardiology, Brescia, Italy
[4]University of Glasgow, School of Psychology & Neuroscience, Glasgow, UK
[5]Vivantes Klinikum, Department of Cardiology and Cardiovascular Clinical Research Unit, Berlin, Germany

Abstract

The Trustworthy-AI Medical Image Analysis group at the University of Brescia is a team dedicated to advancing the field of medical image analysis through collaborative research activities. The group's efforts are concentrated on the development of innovative systems and solutions to address complex image interpretation challenges, specifically within two imaging modalities: Brain MRI and Chest X-ray, and their corresponding anatomical districts.

The group's research efforts are aimed at improving the accuracy, speed, and efficiency of image interpretation, with a focus on ensuring the reliability and safety of AI-assisted medical decision-making processes. By leveraging advanced deep learning techniques, the group aims to develop cutting-edge algorithms that can accurately and efficiently analyze medical images, aiding in the detection, diagnosis, and treatment of various medical conditions.

Keywords

Deep learning, Magnetic Resonance Imaging, Chest X-ray, Brain segmentation, Cortical thickness, COVID-19 prognosis, Cardiovascular risk factors

1. Introduction

The research on deep learning architectures and methods represents the mainstream in the medical image analysis domain, with countless academic contributions and an increasingly relevant market sector in the field of digital healthcare management.

In this report, we summarize some of the activities of our research group in the fields of Brain MRI and Chest X-rays, emphasizing the motivation of the adopted approaches, the main results, and the collaborative nature of the works.

All the described activities have in common the fact that they involve some challenging aspects related to the presence of unmet needs on both new and consolidated diagnostic image interpretation tasks.

On Brain MRI volumes we fight the scanner effect to obtain fast and robust multi-site brain segmentation (Sec.2), presenting preliminary activities tackling some open issues about cortical thickness estimation (Sec.3).

On Chest X-rays (CXR) we present some activities related to COVID-19 prognosis related to our participation in two initiatives: a Best Practice study case in the Z-inspection® framework; and as winners of the explainability track of the AI4COVID Hackathon sponsored by CINI Lab AIIS (Sec.4). Still on CXR, we present our research on new perspectives about the possibility to predict relevant risk factors related to common cardiovascular diseases (Sec.5).

2. Fighting the scanner effect in brain MRI segmentation on multi-site data

Many clinical and research studies of the human brain require accurate structural MRI segmentation. While traditional atlas-based methods can be applied to volumes from any acquisition site, recent deep learning algorithms ensure high accuracy only when tested on data from the same sites exploited in training (*i.e.*, internal data). Performance degradation experienced on external data (*i.e.*, unseen volumes from unseen sites) is due to the inter-site variability in intensity distributions induced by different MR scanner models, acquisition parameters, and unique artefacts. To mitigate this site-dependency, often referred to as the *scanner effect*, we propose `LOD-Brain`, a

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
alberto.signoroni@unibs.it (A. Signoroni); mattia.savardi@unibs.it (M. Savardi)
0000-0002-8383-3766 (A. Signoroni); 0000-0002-2751-5157 (M. Savardi)

3D convolutional neural network with progressive levels-of-detail (LOD), able to segment brain data from any site [1]. Coarser network levels are responsible for learning a robust anatomical prior helpful in identifying brain structures and their locations, while finer levels refine the model to handle site-specific intensity distributions and anatomical variations. We ensure robustness across sites by training the model on an unprecedentedly rich dataset aggregating data from open repositories: almost 27,000 T1w volumes from around 160 acquisition sites, at 1.5 - 3T, from a population spanning from 8 to 90 years old. Extensive tests demonstrate that LOD-Brain produces state-of-the-art results, with no significant difference in performance between internal and external sites, and robust to challenging anatomical variations. Its portability paves the way for large-scale applications across different healthcare institutions, patient populations, and imaging technology manufacturers.

2.1. Methods

We introduce LOD-Brain, a progressive level-of-detail network for training a robust brain MRI segmentation model from a huge variety of multi-site and multi-vendor data. LOD-Brain architecture is organised on multiple levels of detail (LOD), as shown in Fig. 1. Each level is a convolutional neural network (CNN) that processes 3D brain data at a different scale obtained via progressively downsampling the input volume. Thanks to the rich variability of brain samples coming from 70 datasets from different MRI acquisition sites, the proposed architecture learns, at lower levels, a robust brain anatomical prior. Concurrently, higher levels handle site-specific intensity distributions and scanner artefacts. Through inter-level connections between networks and a bottom-up training procedure, such architecture integrates contributions from all levels to produce an accurate and fast segmentation.

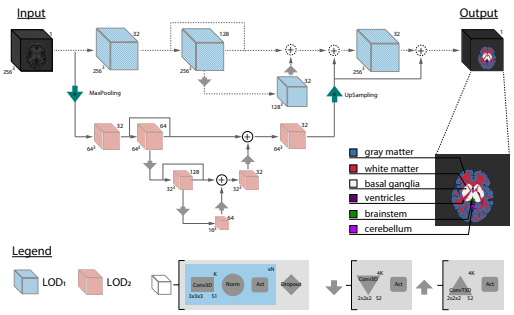


Figure 1: LOD-Brain architecture selected for the experiments on the brain MRI segmentation task. The lower level learns a coarse and site-independent brain representation, while the superior one incorporates the learnt spatial context, and refines segmentation masks at finer scales.

2.2. Results

LOD-Brain shows outstanding generalisation capabilities, as it performs better than other state-of-the-art solutions on almost every novel site, with no need for retraining nor fine-tuning, and with no relevant performance offset in segmenting either internal or external sites. Furthermore, it proves to be general and robust across sites against different population demographics, anatomical challenges, clinical conditions, and technical specifications (*e.g.*, field strength, manufacturer).

As an open-source tool, LOD-Brain can be used off-the-shelf on unseen scans from novel sites. Segmentation masks are returned very quickly (a few seconds on a GPU) thanks to a reduced number of model parameters ($300K$), if compared to other state-of-the-art solutions.

A comparative assessment of our method against state-of-the-art techniques (we use FreeSurfer[2] as silver GT reference) is proposed here in terms of both brain segmentation performance and model complexity. The considered benchmark methods are: QuickNat [3], SynthSeg [4], 3D-UNet [5], CEREBRUM [6], FastSurferCNN [7]. Fig. 2 shows the obtained results on the whole testing set grouped by segmented brain structure. Obtained results highlight LOD-Brain as one of the most competing methods on all brain labels, as it yields the best scores in almost all target structures and on the majority of external datasets with good-quality ground truth labels. The number of parameters for each model is also reported, highlighting LOD-Brain as the best overall model in terms of performance-to-complexity ratio. Many more details about methods, results as well as code, model, and demo are available on the project website. For example, it is relevant to note the high performance achieved on the ABCD dataset, despite it includes volumes from 32 diverse scanners, previously skull-stripped and aligned to MNI152 reference space (a common procedure in this domain).

3. A method for estimating cortical thickness in Brain MRI

Studying brain anatomical deviations from normal progression along the lifespan is essential to understand inter-individual variability and its relation to the onset and progression of several clinical conditions [8]. Among available quantitative measurements, mean *cortical thickness* across the brain has been associated with normal ageing and neurodegenerative conditions like mild cognitive impairment, Alzheimer's disease, frontotemporal

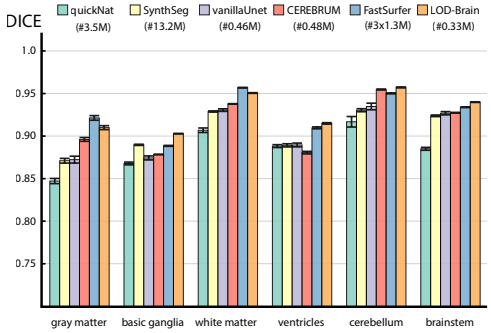


Figure 2: Performance comparison (results grouped by brain structure): QuickNat [3], SynthSeg [4], 3D-UNet [5], CEREBRUM [6], FastSurferCNN [7], and our method. Results are computed on the test set of 5, 956 volumes, using FreeSurfer[2] as GT reference and grouped for brain structure. Number of parameters for each model are reported.

dementia, Parkinson's disease, amyotrophic lateral sclerosis, and vascular cognitive impairment. Automatic techniques, such as FreeSurfer[2] and CAT12 Toolbox [9] offer out-of-the-box cortical thickness estimates, but with an excessively long computational time (up to 10 hours per volume). Moreover, comparison studies have found systematic differences between these approaches [10], with discrepancies particularly pronounced in clinical data [11], questioning the reliability of these CT estimations. As more and more studies in medicine and neuroscience analyse hundreds to thousands of brain MRI scans, there is a growing need for automatic, fast, and reliable tools for cortical thickness estimation.

3.1. Methods

We propose a method for estimating cortical thickness from MRI in just a few seconds [12]. The proposed framework, shown in Figure 3, exploits our recent achievements in deep learning segmentation methods [6, 1] for extracting grey and white matter segmentation masks and the related probability maps from an MRI T1w volume. All these volumes are given as inputs to a Convolutional Neural Network trained to compute both the external grey matter surface (or pial) and the related thickness.

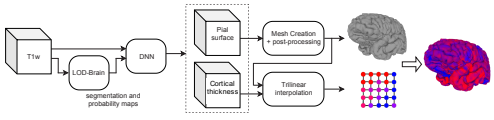
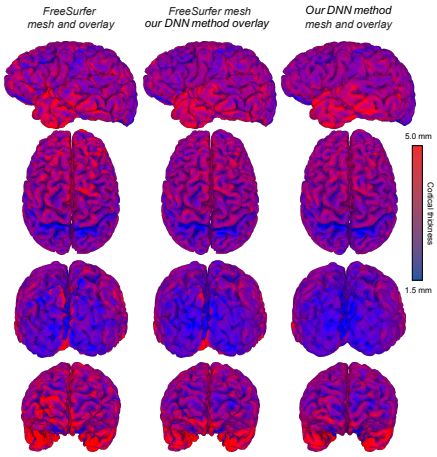
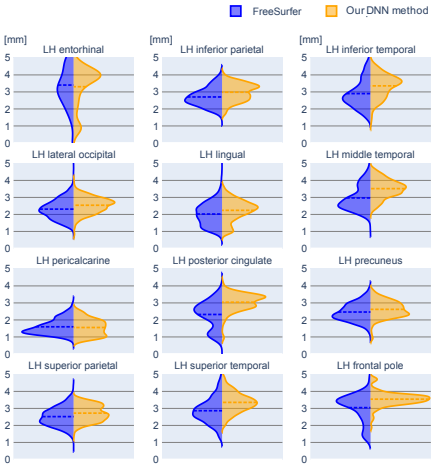


Figure 3: Framework for cortical thickness and surface estimation.



(a) Visual results



(b) Cortical thickness maps comparison

Figure 4: (a) - Visual results of FreeSurfer mesh and CT overlay, FreeSurfer mesh and our DNN method overlay, and our DNN method mesh and overlay. (b) - Comparison of the distributions of the cortical thickness values of 12 left hemisphere regions for FreeSurfer (blue) and our DNN method (orange) on one testing subject in mm. Dotted lines represent average values; higher symmetry in distributions denotes higher region-wise *cortical thickness* similarity. Similar results are obtained for the right hemisphere and other subjects.

The supervised model is trained, with volumes obtained by FreeSurfer [2] as ground truth. The network architecture resembles a 3D U-Net, with 4 levels of convolutional layers, and two output branches predicting the pial surface and the cortical thickness. Training, validation, and testing volumes are obtained from the AOMIC dataset, counting 1311, 100, and 500 volumes respectively.

3.2. Results

In Figure 4-(a), we show qualitative results highlighting how our method performs with respect to FreeSurfer, in both the mesh generation and the cortical thickness estimation. In Figure 4-(b), we compare numerically the cortical thickness estimation distributions obtained with FreeSurfer and our method on a testing subject. Our DNN method [12] is the first DL-based approach for cortical thickness estimation on structural MRI. The extraction of cortical thickness distributions in just a few seconds unlocks the ability to quickly draw population trajectories for thousands of healthy subjects' data, creating an atlas with different distributions for different brain areas.

4. Trustworthy AI for COVID-19 severity estimation and prognosis

Although during the COVID-19 pandemic, the AI-based interpretation of CXRs focused largely on COVID-19 diagnosis, few studies addressed other relevant tasks such as severity estimation, deterioration, and prognosis also trying to explain the models' decisions. A recent international hackathon sponsored by CINI Lab AIIS during the Dubai Expo 2020 sought to develop machine learning (ML) models to predict COVID-19 prognosis and explain their predictions in a clinically interpretable manner. The hackathon dataset included CXRs and clinical features collected during triage for a large number of subjects. To calculate the prognostic value, a deep learning model estimated the lung compromise degree from the CXRs, which was considered alongside the clinical features. Then, we trained and evaluated multiple models to identify the best-performing, fine-tuning them before inference and generating visual and numerical explanations to justify their predictions. Our model achieved high accuracy, ranking second in the final rankings with 75% and 73.9% in sensitivity and specificity. In terms of explainability, it was agreed to be the most interpretable by health professionals and was ranked first. Our study [13] highlights the potential of ML models in helping physicians formulate trustworthy COVID-19 prognoses, contributing to the efforts to improve the allocation of limited healthcare resources.

4.1. Methods

The dataset included a blind test set and a training set with more than 1100 subjects, characterized by 38 clinical features and a CXR image. After imputing missing values in the former and improving the quality of the latter, we exploited BSNet [14] to predict the multi-regional lung compromise index Brixia-score [15] for each training subject from its CXR. A posthoc trustworthy assessment, called Z-Inspection® [16], was applied to this network and its deployment in the radiology department of the ASST Spedali Civili clinic in Brescia, Italy during the pandemic time. The predicted Brixia-score and other parameters were found as clinically significant by a model-based feature extraction procedure and constituted the feature set on which multiple models were trained on. Once identified the best-performing on an internal validation set, we employed it to predict the prognosis for the subjects in the test set. Finally, we produced both visual and numerical explanations to justify the model's predictions from both a global and a patient-specific perspective.

4.2. Results

The best-performing model was a Random Forest (RF). The RF proved to be accurate on the test set and was ranked second in the final rankings with 75% and 73.9% in sensitivity and specificity, respectively. From a global perspective, the most important features to our RF to make its decisions were blood pressure, Brixia-score, and LDH enzyme concentration. Conversely, from a patient-specific perspective, we used SHapley Additive exPlanations (SHAP [17]) values-based charts to justify the RF's predictions. Such charts, of which an example is depicted in Fig. 5, show which clinical features pushed the RF to predict a certain prognosis, how "strongly", and which ones pushed it to predict the opposite prognosis.

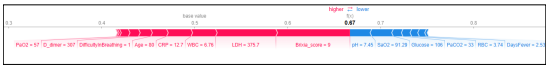


Figure 5: SHAP values-based chart showing patient-specific features (in red) that drove the RF to predict a severe prognosis against other features (in blue) that pushed it to predict a mild prognosis. The final prediction was correctly severe.

Finally, the last patient-specific explanation, shown in Fig. 6, was provided by the explainability map produced by BSNet highlighting which regions of the lungs contributed most to which local severity score.
All these explanations were agreed to be highly interpretable by a panel of health professionals and radiologists. For this reason, our model was ranked first in the final clinical explainability ranking.

5. AI to predict cardiovascular risk factors from Chest X-rays

Coronary artery disease (CAD) is the single leading cause of mortality, premature death, and morbidity worldwide. Artificial intelligence (AI) could help identify markers present within first-line diagnostic imaging routinely performed in patients referred for suspected angina, such as

Figure 6: From left to right, a patient's pre-processed CXR, its predicted Brixia-score and the corresponding super-pixel-based explainability map of the lungs. The higher a super-pixel's colour saturation, the greater that superpixel's contribution to the regional severity score associated with the same colour.

CXRs. The objective of our work is to train, test, and clinically validate deep learning (DL) algorithms for detecting the presence of significant CAD based on CXRs [18]. The CXR modality is ubiquitous and carries a plethora of information concerning the patient's health status, including direct and indirect signs of CAD. Our DL algorithm can predict, with high sensitivity, the presence of severe CAD in patients referred for suspected angina. It could be used to pre-test significant CAD probability in outpatient clinics, emergency room settings, and CAD screening in more extensive settings. Further studies are required to externally validate the algorithm and develop a clinically applicable tool.

5.1. Methods

Data from patients undergoing chest radiography and coronary angiography were retrospectively analysed. A deep convolutional neural network (DCNN) was designed to detect significant CAD from the patient posteroanterior/anteroposterior chest radiograph. The DCNN was trained for binary classification of severe CAD absence/presence (at least one diseased coronary vessel with ≥ 70% stenosis). Coronary angiography reports were used as the ground truth. Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) of the DCNN were calculated. Multivariate analysis was performed to identify independent correlation among the presence of significant CAD (dependent variable), DCNN prediction, and CAD risk factors.

5.2. Results

Information of 7728 patients referred for suspected angina was reviewed. Severe CAD was present in 4482 patients (58%; 1% left main, 28% one vessel, 16% two vessels, and 12% 3 vessels). Patients were randomly divided for training (70%; n = 5454) and fine-tuning/testing (10%; n = 773) of the algorithm. Internal validation was performed with the remaining patients (20%; n = 1501). The DCNN

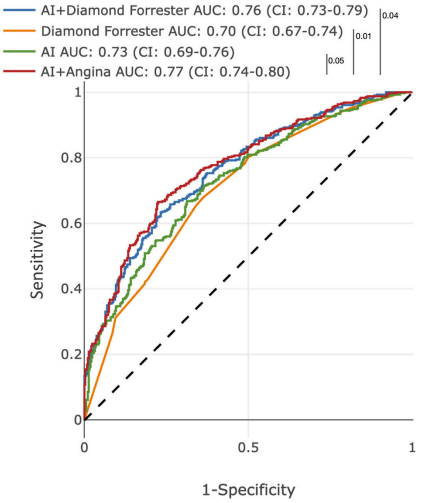


Figure 7: Receiver operating characteristic (ROC) curves for the binary classification of the presence significant coronary artery disease (CAD) demonstrating area under the curve (AUCs) of 0.77 for AI + angina Type; 0.76 for AI + Diamond Forrester; 0.73 for AI and 0.70 for Diamond Forrester.

had a processing time of about 20 s to analyse the whole internal validation set on a workstation with an Nvidia Titan V Graphics Processing Unit (GPU) with 12GB of memory. At binary logistic regression, the DCNN prediction was the strongest independent determinant of severe CAD (p<0.0001; OR: 50.7; CI: 24.0-107.0). Age (p=0.006; OR: 1.01; CI: 1.0-1.02) and Diamond-Forrester score (p<0.0001; OR: 1.022; CI: 1.018-1.026) were also independently related to CAD, although with lower significance and odds-ratios. Using an operating cut-point with high sensitivity, the DCNN had a sensitivity of 0.90 and specificity of 0.31 to detect significant CAD in the internal validation group (AUC 0.73; 95% CI DeLong, 0.69-0.76). Adding to the AI chest radiograph interpretation, patient age and angina status improved the prediction (AUC 0.77; 95% CI DeLong, 0.74-0.80). ROC curves for the binary CAD classification are reported in Fig.7. Attention maps were created considering the DCNN with the highest performance. Heat map activations are primarily localized to the cardiac silhouette, left ventricular apex, pulmonary bases, pulmonary parenchyma, costophrenic sinuses, pulmonary hila, thoracic aorta, supra-aortic vessels, and clavicle region (Fig.8, panels A-F).

References

[1] M. Svanera, M. Savardi, A. Signoroni, S. Benini, and L. Muckli. *Fighting the scanner effect in brain MRI segmentation with a progressive level-of-detail*

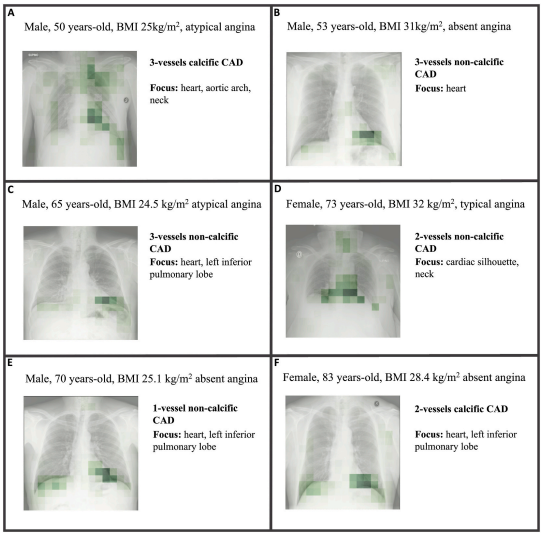


Figure 8: (A-F). Heat maps of 6 patients affected by severe CAD. Areas suggestive of CAD presence are highlighted in degrees of green tonality.

network trained on multi-site data. 2022. arXiv: 2211.02400 [eess.IV].

[2] B. Fischl. "FreeSurfer". In: *NeuroImage* 62.2 (2012), pp. 774–781.

[3] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, and ADNI. "QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy". In: *NeuroImage* 186 (2019), pp. 713–727.

[4] B. Billot et al. "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". In: *Medical Image Analysis* 86 (2023), p. 102789.

[5] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Springer, 2016, pp. 424–432.

[6] D. Bontempi, S. Benini, A. Signoroni, M. Svanera, and L. Muckli. "CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI". In: *Medical Image Analysis* 62 (2020).

[7] L. Henschel et al. "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219 (2020), p. 117012.

[8] R. A. Bethlehem et al. "Brain charts for the human lifespan". In: *Nature* 604.7906 (2022), pp. 525–533.

[9] J. Ashburner. "SPM: a history". In: *Neuroimage* 62.2 (2012), pp. 791–800.

[10] R. Seiger, S. Ganger, G. S. Kranz, A. Hahn, and R. Lanzenberger. "Cortical thickness estimations of FreeSurfer and the CAT12 toolbox in patients with Alzheimer's disease and healthy controls". In: *Journal of Neuroimaging* 28.5 (2018), pp. 515–523.

[11] M. Ozzoude et al. "Cortical thickness estimation in individuals with cerebral small vessel disease, focal atrophy, and chronic stroke lesions". In: *Frontiers in Neuroscience* 14 (2020), p. 598868.

[12] D. Ferrari et al. "A Deep Learning Method for Brain MRI Cortical Thickness Estimation". In: *OHBM 2023 Annual Meeting*. accepted presentation. 2023.

[13] E. Coppola, D. Ferrari, M. Savardi, and A. Signoroni. "Explainable AI for COVID-19 prognosis from early Chest X-ray and clinical data in the context of the COVID-CXR international hackathon". In: *Proceedings of SPIE Medical Imaging 2023: Computer-Aided Diagnosis*. Vol. SPIE 12465. in press. 2023.

[14] A. Signoroni et al. "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset". In: *Medical Image Analysis* 71 (2021), p. 102046.

[15] A. Borghesi and R. Maroldi. "COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression". In: *La radiologia medica* 125.5 (2020), pp. 509–513.

[16] H. Allahabadi et al. "Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients". In: *IEEE Transactions on Technology and Society* 3.4 (2022), pp. 272–289.

[17] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *NIPS*. 2017, pp. 4765–4774.

[18] G. D'Ancona et al. "Deep learning to detect significant coronary artery disease from plain chest radiographs AI4CAD". In: *International Journal of Cardiology* 370 (2023), pp. 435–441.