Overview of ImageCLEFtuberculosis 2019 — Automatic CT–based Report Generation and Tuberculosis Severity Assessment

Yashin Dicente Cid[1], Vitali Liauchuk[2], Dzmitri Klimuk[3], Aleh Tarasau[3], Vassili Kovalev[2], and Henning Müller[1,4]

[1] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
[2] United Institute of Informatics Problems, Minsk, Belarus;
[3] Republican Research and Practical Centre for Pulmonology and TB, Minsk, Belarus;
[4] University of Geneva, Switzerland
yashin.dicente@hevs.ch

Abstract. ImageCLEF is the image retrieval task of the Conference and Labs of the Evaluation Forum (CLEF). ImageCLEF has historically focused on the multimodal and language-independent retrieval of images. Many tasks are related to image classification and the annotation of image data as well as the retrieval of images. Since 2017, when the tuberculosis task started in ImageCLEF, the number of participants has kept growing. In 2019, 13 groups from 11 countries participated in at least one of the two subtasks proposed: (1) SVR subtask: the assessment of a tuberculosis severity score and (2) CTR subtask: the automatic generation of a CT report based on six relevant CT findings. In this second edition of the SVR subtask the results support the assessment of a severity score based on the CT scan with up to 0.79 area under the curve (AUC) and 74% accuracy, so very good results. In addition, in the first edition of the CTR subtask, impressive results were obtained with 0.80 average AUC and 0.69 minimum AUC for the six CT findings proposed.

Keywords: Tuberculosis, Computed Tomography, Image Classification, Severity Scoring, Automatic Reporting, 3D Data Analysis

1 Introduction

ImageCLEF[5] is the image retrieval task of CLEF (Conference and Labs of the Evaluation Forum). ImageCLEF was first held in 2003 and in 2004 a medical task was added that has been held every year since then [1–4]. More information on the other tasks organized in 2019 can be found in [5] and the past editions are described in [6–10].

[5] http://www.imageclef.org/

Tuberculosis (TB) is a bacterial infection caused by a germ called Mycobacterium tuberculosis. About 130 years after its discovery, the disease remains a persistent threat and a leading cause of death worldwide [11]. This bacterium usually attacks the lungs but it can also damage other parts of the body. Generally, TB can be cured with antibiotics. However, the greatest problem that can happen to a patient with TB is that the organisms become resistant to two or more of the standard drugs. In contrast to drug sensitive (DS) TB, its multi-drug resistant (MDR) form is much more difficult and expensive to recover from. Thus, early detection of the MDR status is fundamental for an effective treatment. The most commonly used methods for MDR detection are either expensive or take too much time (up to several months) to really help in this scenario. Therefore, there is a need for quick and at the same time cheap methods of MDR detection. In 2017, ImageCLEF organized the first challenge based on Computed Tomography (CT) image analysis of TB patients [12], with a dedicated subtask for the detection of MDR cases. The classification of TB subtypes was also proposed in 2017. This is another important task for TB analysis since different types of TB should be treated in different ways. Both subtasks were also proposed in the 2018 edition where we extended their respective data sets. Moreover, a new subtask was added based on assessing a severity score of the disease given a CT image.

This article first describes the two subtasks proposed around TB in 2019. Then, the data sets, evaluation methodology and participation are detailed. The results section describes the submitted runs and the results obtained for the two subtasks. A discussion and conclusion section ends the paper.

2 Tasks, Data Sets, Evaluation, Participation

2.1 The Tasks in 2019

Two subtasks were organized in 2019, one was common with the 2018 edition and one new subtask was added:

- Severity score assessment (SVR subtask).
- Automatic CT report generation (CTR subtask).

This section gives an overview of each of the two subtasks.

SVR - Severity Scoring: As in 2018, the goal of this subtask is to assess the severity based on the CT image and additional clinically relevant meta-data. The severity score is a cumulative score of severity of a TB case assigned by a medical doctor. Originally, the score varied from 1 ("critical/very bad") to 5 ("very good"). The original severity score was included as training meta-data but the final score that participants had to assess was reduced to a binary category: "LOW" (scores 4 and 5) and "HIGH" (scores 1, 2 and 3).

CTR - CT Report: In this subtask the participants had to generate an automatic report based on the CT image. This report had to include the following CT findings in binary form (0 or 1): Left lung affected, right lung affected, lung capacity decrease, presence of calcifications, presence of pleurisy and presence of caverns.

2.2 Data Sets

In 2019, both subtasks (SVR and CTR) used the same data set containing 335 chest CT scans of TB patients along with a set of clinically relevant meta-data, divided into 218 patients for training and 117 for testing. The selected meta-data include the following binary measures: disability, relapse, symptoms of TB, comorbidity, bacillary, drug resistance, higher education, ex-prisoner, alcoholic, smoking, and severity. Table 1 details the distribution of patients within each label for the SVR and CTR subtasks.

Table 1. Distribution of patients within each label for the SVR and CTR subtasks.

Set	High severity	Left lung affected	Right lung affected	Lung capacity decrease	Pres. of calcif.	Pres. of pleurisy	Pres. of caverns
Train	107 (49%)	156 (72%)	177 (81%)	64 (29%)	28 (13%)	16 (7%)	89 (41%)
Test	55 (47%)	74 (63%)	95 (81%)	8 (7%)	60 (51%)	10 (9%)	40 (34%)

For all patients we provided 3D CT images with a slice size of 512×512 pixels and number of slices varying from about 50 to 400. All the CT images were stored in NIFTI file format with .nii.gz file extension (g-zipped .nii files). This file format stores raw voxel intensities in Hounsfield units (HU) as well the corresponding image meta-data such as image dimensions, voxel size in physical units, slice thickness, etc. The entire dataset including the CT images and the associated meta-data were provided by the Republican Research and Practical Center for Pulmonology and Tuberculosis that is located in Minsk, Belarus. The data were collected in the framework of several projects that aim at the creation of information resources on the lung TB and drug resistance challenges. The projects were conducted by a multi-disciplinary team and funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), U.S. Department of Health and Human Services, USA, through the Civilian Research and Development Foundation (CRDF). The dedicated web-portal[6] developed in the framework of the projects stores information of more than 940 TB patients from five countries: Azerbaijan, Belarus, Georgia, Moldova and Romania. The information includes CT scans, X-ray images, genome data, clinical and social data.

[6] `http://tbportals.niaid.nih.gov/`

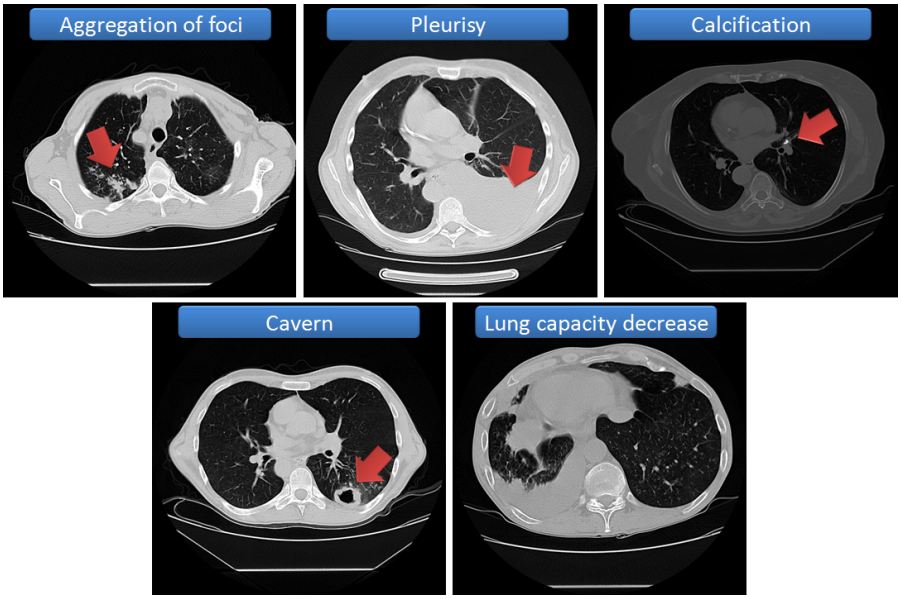

Fig. 1. Slices of typical CT images with several types of TB-related findings.

For all patients we provided automatically extracted masks of the lungs obtained using the method described in [13]. The masks were manually analyzed based on statistics on number of lungs found and size ratio between right and left lung. Only the masks with anomalies on these statistics were visualized. The code used to segment the patients was adapted for the cases with unsatisfactory segmentation. After this, all patients with anomalies presented a satisfactory mask.

Pathological changes in lungs affected by tuberculosis may be represented by a large variety of findings. In most cases such finding include aggregations of foci and infiltrations of different sizes. However, rarer types of lesions may be present including fibrosis, atelectasis, pneumathorax, etc. "Left lung affected" and "Right lung affected" labels provided with the CTR data set indicated presence of any kind of TB-associated lesions in the left and right lung, respectively. Typical examples of CT findings are shown in Fig. 1. Pleurisy, calcifications, caverns and lung capacity decrease were considered separately from the other types of lesions. Pleurisy is known as inflammation of the membranes that surround the lungs and line the chest cavity[7]. Calcifications are usually represented by densely calficied foci that look like bright spots (usually more than 1000 Hounsfield Units) on CT images [14]. Calcifications may occur inside of lungs but also can be located on vessels and the mediastinum. Caverns, also known as pulmonary cavities, are gas-filled areas of the lung in the center of nodules or areas of consolidation [15]. Lung capacity decreased indicates the decrease of volume of the affected lungs compared to normal lungs. Lung capacity de-

[7] `https://www.nhlbi.nih.gov/health-topics/pleurisy-and-other-pleural-disorders`

Table 2. List of participants submitting a run to at least one subtask.

Group name	Main institution	Country	Subtask SVR	Subtask CTR
CompElecEngCU	Çukurova University	Turkey	×	×
FIIAugt	Alexandru Ioan Cuza University of Iaşi	Romania	×	
HHU	Heinrich Heine University	Germany	×	×
LIST	Abdelmalek Essaâdi University	Morocco		×
MedGIFT	University of Applied Sciences Western Switzerland (HES–SO)	Switzerland	×	×
MostaganemFSEI	University of Abdelhamid Ibn Badis Mostaganem	Algeria	×	×
PwC	PwC	India		×
SD VA HCS/UCSD	San Diego VA Health Care System	USA	×	×
SSN CoE	SSN College of Engineering	India	×	
UIIP	United Institute of Informatics Problems	Belarus	×	×
UIIP_BioMed	United Institute of Informatics Problems	Belarus	×	×
UniversityAlicante	University of Alicante	Spain	×	×
UoAP	University of Asia Pacific	Bangladesh	×	

creased can be caused by many factors and can be often associated with other CT findings such as pleurisy and presence of large caverns.

2.3 Evaluation Measures and Scenario

Similar to the previous editions, the participants were allowed to submit up to 10 runs to each of the two subtasks. In the case of the SVR task, the participants had to provide the probability of HIGH severity for each patient. During the challenge, this task was evaluated with area under the receiver operating characteristic (ROC) curve (AUC) and accuracy and the runs were ranked first by AUC and then by accuracy. Moreover, we included the unbalanced Cohen kappa coefficient to our analysis and the ROC curves are provided in Section 3.

In the case of the CTR task, the participants had to provide the probability of each CT finding (see Section 2.1) for each patient, i.e. for each patient they had to provide a 6-dimensional vector with the probabilities. This task was considered a multi-binary classification problem and standard binary classification metrics are provided. During the challenge the runs were ranked based on the average AUC and the min AUC obtained. In addition, since the data set was highly unbalanced for some of the CT findings (see Table 1), we include the AUC, sensitivity and specificity for each finding.

2.4 Participation

In 2019 there were 97 registered teams and 48 signed the end user agreement. 13 groups from 11 countries participated in one or more subtasks and submitted

results. These numbers are similar to 2017 and 2018, where there were ∼90 registered teams, ∼50 that signed the end user agreement, and ∼10 teams from 9 countries submitting results. Table 2 shows the list of participants and the subtasks where they participated.

3 Results

This section provides the results obtained by the participants in each of the subtasks.

3.1 SVR Subtask

Table 3 shows the AUC and accuracy obtained by each participant's run, measures used to establish the SVR ranking. The ROC curve for the best run of each participant is shown in Figure 2. In addition, Table 4 summarizes the results for each best run and includes the unweighted Cohen Kappa coefficient. The best

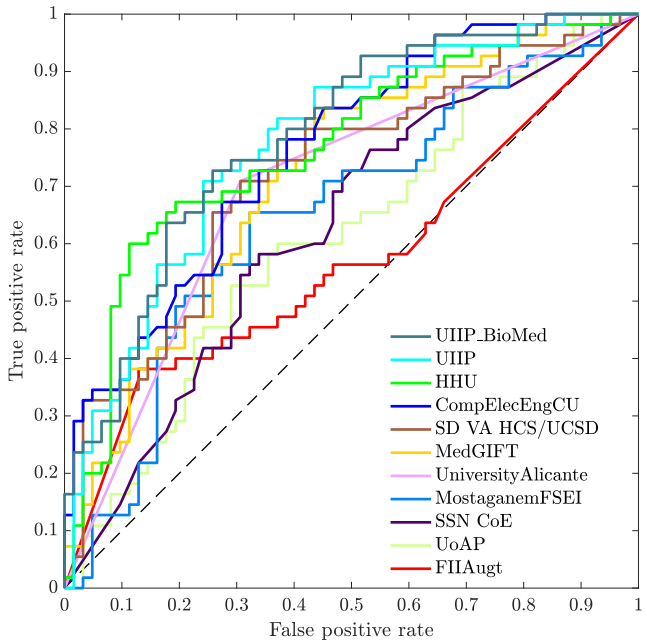

Fig. 2. Receiver operating characteristic (ROC) curves obtained by the best run of each group. The dashed line marks the curve of a random classifier.

results were obtained by the UIIP_BioMed [16] group, both in terms of AUC and

Table 3. Results obtained by the participants in the SVR subtask.

Group name	Run	AUC	Accuracy	Rank
UIIP_BioMed	SRV_run1_linear.txt	**0.7877**	0.7179	1
UIIP	subm_SVR_Severity	0.7754	0.7179	2
HHU	SVR_HHU_DBS2_run01.txt	0.7695	0.6923	3
HHU	SVR_HHU_DBS2_run02.txt	0.7660	0.6838	4
UIIP_BioMed	SRV_run2_less_features.txt	0.7636	**0.7350**	5
CompElecEngCU	SVR_mlp-text.txt	0.7629	0.6581	6
SD VA HCS/UCSD	SVR_From_Meta_Report1c.csv	0.7214	0.6838	7
SD VA HCS/UCSD	SVR_From_Meta_Report1c.csv	0.7214	0.6838	8
MedGIFT	SVR_SVM.txt	0.7196	0.6410	9
SD VA HCS/UCSD	SVR_Meta_Ensemble.txt	0.7123	0.6667	10
SD VA HCS/UCSD	SVR_LAstEnsembleOfEnsemblesReportCl.csv	0.7038	0.6581	11
UniversityAlicante	SVR-SVM-axis-mode-4.txt	0.7013	0.7009	12
UniversityAlicante	SVR-SVM-axis-mode-8.txt	0.7013	0.7009	13
UniversityAlicante	SVR-MC-4.txt	0.7003	0.7009	14
UniversityAlicante	SVR-MC-8.txt	0.7003	0.7009	15
SD VA HCS/UCSD	SVRMetadataNN1_UTF8.txt	0.6956	0.6325	16
UIIP	subm_SVR_Severity	0.6941	0.6496	17
UniversityAlicante	SVR-LDA-axis-mode-4.txt	0.6842	0.6838	18
UniversityAlicante	SVR-LDA-axis-mode-8.txt	0.6842	0.6838	19
UniversityAlicante	SVR-SVM-axis-svm-4.txt	0.6761	0.6752	20
UniversityAlicante	SVR-SVM-axis-svm-8.txt	0.6761	0.6752	21
MostaganemFSEI	SVR_FSEI_run3_resnet_50_55.csv	0.6510	0.6154	22
UniversityAlicante	SVR-LDA-axis-svm-4.txt	0.6499	0.6496	23
UniversityAlicante	SVR-LDA-axis-svm-8.txt	0.6499	0.6496	24
MostaganemFSEI	SVR_run8_lstm_5_55_sD_lungnet.csv	0.6475	0.6068	25
MedGIFT	SVR_GNN_nodeCentralFeats_sc.csv	0.6457	0.6239	26
HHU	run_6.csv	0.6393	0.5812	27
SD VA HCS/UCSD	SVT_Wisdom.txt	0.6270	0.6581	28
SSN CoE	SVRtest-model1.txt	0.6264	0.6068	29
HHU	run_8.csv	0.6258	0.6068	30
SSN CoE	SVRtest-model2.txt	0.6133	0.5385	31
UoAP	SVRfree-text.txt	0.6111	0.6154	32
MostaganemFSEI	SVR_FSEI_run2_lungnet_train80_10slices.csv	0.6103	0.5983	33
HHU	run_4.csv	0.6070	0.5641	34
SSN CoE	SVRtest-model3.txt	0.6067	0.5726	35
HHU	run_7.csv	0.6050	0.5556	36
UoAP	SVRfree-text.txt	0.5704	0.5385	37
FIIAugt	SVRab.txt	0.5692	0.5556	38
HHU	run_3.csv	0.5692	0.5385	39
MostaganemFSEI	SVR_FSEI_run6_fuson_resnet_lungnet_10slices.csv	0.5677	0.5128	40
MedGIFT	SVR_GNN_node2vec.csv	0.5496	0.5726	41
MedGIFT	SVR_GNN_nodeCentralFeats.csv	0.5496	0.4701	42
SSN CoE	SVRtest-model4.txt	0.5446	0.5299	43
HHU	run_5.csv	0.5419	0.5470	44
HHU	SVRbaseline_txt.txt	0.5103	0.4872	45
MostaganemFSEI	SVR_FSEI_run4_semDesc_SVM_10slices.csv	0.5029	0.5043	46
MostaganemFSEI	SVR_run7_inception_resnet_v2_small_54_[...].csv	0.4933	0.4701	48
MedGIFT	SVR_GNN_node2vec_pca.csv	0.4933	0.4615	47
MostaganemFSEI	SVR_FSEI_run5_contextDesc_RF_10slices.csv	0.4783	0.4957	49
MostaganemFSEI	SVR_fsei_run0_resnet50_modelA.csv	0.4698	0.4957	50
MostaganemFSEI	SVR_FSEI_run9_oneSVM_desSem_10slices_[...].csv	0.4636	0.5214	51
HHU	run_2.csv	0.4452	0.4530	52
MedGIFT	SVR_GNN_node2vec_pca_sc.csv	0.4076	0.4274	53
MostaganemFSEI	SVR_FSEI_run10_RandomForest_semDesc_[...].csv	0.3475	0.4615	54

Table 4. Detailed results obtained in the SVR task by the best run of each group.

Group name	AUC	Accuracy	Kappa
UIIP_BioMed	**0.7877**	**0.7179**	0.4310
UIIP	0.7754	**0.7179**	**0.4321**
HHU	0.7695	0.6923	0.3862
CompElecEngCU	0.7629	0.6581	0.3289
SD VA HCS/UCSD	0.7214	0.6838	0.3646
MedGIFT	0.7196	0.6410	0.2720
UniversityAlicante	0.7013	0.7009	0.4014
MostaganemFSEI	0.6510	0.6154	0.2335
SSN CoE	0.6264	0.6068	0.2109
UoAP	0.6111	0.6154	0.2272
FIIAugt	0.5692	0.5556	0.1005

accuracy. The same group also ranked first in the previous edition, obtaining a significant improvement this year: from 0.7025 to 0.7877 AUC. For this edition, they proposed an initial convolutional neural network (CNN) using 2D projections of the 3D CT scans that provides a probability of high TB severity. Then they combined these probabilities with the available meta-data and used a linear regression classifier to provide the final classification score. The UIIP [17] group obtained the best Kappa. In their approach, they first performed data augmentation and used a 3D CNN as autoencoder, followed by a traditional classifier, such as random forest.

A total of five groups (including UIIP_BioMed) participated in both editions of this subtask (2018 and 2019), and all obtained higher results in 2019: HHU [18] improved from 0.6484 to 0.7695 AUC. They proposed a completely new approach where they first assessed the CT-findings proposed in the CTR subtask and then applied linear regression to obtain the severity score. In addition, they also tried a different approach based on selecting 16 CT slices and using a 3D CNN (UNet) that obtained lower results. The SD VA HCS/UCSD [19] used an ensemble of 2D CNNs, combining the predictive scores provided by each CNN. Then they fuse these scores with the meta-data into a Support Vectors Machine (SVM) classifier that provided the final severity score. With this approach they went from 0.6658 to 0.7214 AUC. The performance of MedGIFT [20] remained almost the same between both editions (0.7162 vs 0.7196 AUC). Their best approach in 2019 is similar to the one proposed in 2018. They proposed to model the lung as a graph by dividing the lung fields into a number of subregions (different for each patient) and considering these subregions as nodes of a graph. They then defined weighted edges between adjacent subregions, where the weights encode the distance between 3D texture descriptors obtained in each subregion (node). In order to compare the obtained graphs, they transform these graphs into a lung descriptor vector and used SVM to classify them. In addition, they also attempted to classify the graphs with a 2D CNN obtaining much lower results. Finally, the last group participating in both editions is the MostaganemFSEI [21] group, that improved from 0.5987 to 0.6510 AUC. Their pipeline consisted on

first selecting meaningful axial CT slices manually. These slices are then described with semantic features extracted via a 2D CNN. As a last step, they used a 5-class long short term memory (LSTM) algorithm to obtain one of the original 5 levels of TB severity that is then transformed into the classes high or low.

CompElecEngCU [22] created 2D derived images by concatenating sagittal and coronal CT slices that are classified with a hybrid of a 2D CNN based on AlexNet and a Multi-Layer Perceptron. The UniversityAlicante [23] group considered each CT volume as a time series (or video) and used optical flow on the 3 directions. The SSN CoE [24] and UoAP [25] groups used a similar approach. Both first manually selected a set of relevant slices for each patient and then used a CNN. In the case of SSN CoE they selected 30 slices and used a 2D CNN. UoAP used a 3D CNN (VoxNet) with either 16 or 32 CT slices. Finally, the FIIAugt [26] group performed random sampling of pixels of the CT volumes and used a combination of decision trees and weak classifiers.

3.2 CTR Subtask

To provide a ranking in this subtask we used the mean AUC and min AUC over the six binary CT-findings proposed. Table 5 provides these two measures for all runs submitted. Similar to the SVR subtask we provide more detailed results for the best run of each group. For each best run and for each CT-finding, Figures 3, 4, 5 and 6 depicts the ROC curves, AUC, sensitivity and specificity, respectively. In this case, the sensitivity and specificity metrics have been computed assuming the standard decision threshold of 0.50. Moreover, Table 6 summarizes the results of the best runs providing mean, min and max values for each of these metrics.

Again, UIIP_BioMed [16] is the winner of this subtask with a mean AUC of 0.7968 and a min AUC of 0.6860. When we check the individual AUCs for each CT-finding (see Figure 4), we observe that they outperformed every other method in the left and right lung labels by a high margin. However, they have similar results to other techniques in the other four CT-findings. In this subtask they used different approaches for each CT-finding, mainly consisting of a unique 2D CNN architecture with modified input for each abnormality. It is worth to mention their simple technique for detecting pleurisity: they noticed that most of the lung masks provided by the organizers did not contain the areas of the lungs presenting pleurisity. Therefore, they used their own lung segmentation algorithm based on atlas registration. The final score for pleurisity was then computed based on the difference between their masks and the organizer's masks. HHU [18] is the other group that used a specific method for each CT-finding, mainly based on morphological operations and binarizations with a standard classifier as a last step. In the case of the MostaganemFSEI [21] modified the last step of the pipeline applied in the SVR subtask, substituting the LSTM step with an SVM classifier. PwC [27] and LIST only participated in this subtask. The latter did not provide details of their approach. In the case of the PwC group they used 3D CNN with 20 slices for feature extraction and used them along with the meta-data in a random forest classifier. All the other groups participating in

Table 5. Results obtained by the participants in the CTR subtask.

Group Name	Run	Mean AUC	Min AUC	Rank
UIIP_BioMed	CTR_run3_pleurisy_as_SegmDiff.txt	**0.7968**	**0.6860**	1
UIIP_BioMed	CTR_run2_2binary.txt	0.7953	0.6766	2
UIIP_BioMed	CTR_run1_multilabel.txt	0.7812	0.6766	3
CompElecEngCU	CTRcnn.txt	0.7066	0.5739	4
MedGIFT	CTR_SVM.txt	0.6795	0.5626	5
SD VA HCS/UCSD	CTR_Cor_32_montage.txt	0.6631	0.5541	6
HHU	CTR_HHU_DBS2_run01.txt	0.6591	0.5159	7
HHU	CTR_HHU_DBS2_run02.txt	0.6560	0.5159	8
SD VA HCS/UCSD	CTR_ReportsubmissionEnsemble2.csv	0.6532	0.5904	9
UIIP	subm_CT_Report	0.6464	0.4099	10
HHU	CTR_HHU_DBS2_run03.txt	0.6429	0.4187	11
HHU	CTR_run_1.csv	0.6315	0.5161	12
HHU	CTR_run_2.csv	0.6315	0.5161	13
MostaganemFSEI	CTR_FSEI_run1_lungnet_50_10slices.csv	0.6273	0.4877	14
UniversityAlicante	svm_axis_svm.txt	0.6190	0.5366	15
UniversityAlicante	mc.txt	0.6104	0.5250	16
MostaganemFSEI	CTR_FSEI_lungNetA_54slices_70.csv	0.6061	0.4471	17
UniversityAlicante	svm_axis_mode.txt	0.6043	0.5340	18
PwC	CTR_results_meta.txt	0.6002	0.4724	19
UniversityAlicante	lda_axis_mode.txt	0.5975	0.4860	20
SD VA HCS/UCSD	TB_ReportsubmissionLimited1.csv	0.5811	0.4111	21
UniversityAlicante	lda_axis_svm.txt	0.5787	0.4851	22
HHU	CTR_run_3.txt.csv	0.5610	0.4477	23
PwC	CTR_results.txt	0.5543	0.4275	24
LIST	predictionCTReportSVC.txt	0.5523	0.4317	25
LIST	predictionModelSimple.txt	0.5510	0.4709	26
MedGIFT	CTR_GNN_nodeCentralFeats_sc.csv	0.5381	0.4299	27
LIST	predictionCTReportLinearSVC.txt	0.5321	0.4672	28
MedGIFT	CTR_GNN_node2vec_pca_sc.csv	0.5261	0.4435	29
LIST	predictionModelAugmented.txt	0.5228	0.4086	30
MedGIFT	CTR_GNN_nodeCentralFeats.csv	0.5104	0.4140	31
MostaganemFSEI	CTR_FSEI_run5_SVM_semDesc_10slices.csv	0.5064	0.4134	32
MedGIFT	CTR_GNN_node2vec_pca.csv	0.5016	0.2546	33
MostaganemFSEI	CTR_FSEI_run4_SVMone_semDesc_[...].csv	0.4937	0.4461	34
MostaganemFSEI	CTR_FSEI_run3_SVMone_semDesc_[...].csv	0.4877	0.3897	35

this subtask, *i.e.* CompElecEngCU, MedGIFT, SD VA HCS/UCSD, UIIP and UniversityAlicante, used the same approach (or with minor modifications) than in the SVR subtask.

4 Discussion and Conclusions

In the second edition of the SVR subtask, we observe a significant improvement by most of the groups that participated in both editions. However, since we transformed the original 5-class regression task into a binary classification problem, AUC is the only metric that we can compare between editions. The final results, around 0.80 AUC and 0.70 accuracy, encourage us to continuing investigating the task. Most of the participants used the clinical meta-data provided, but unfortunately we cannot analyze the individual contribution of these data.

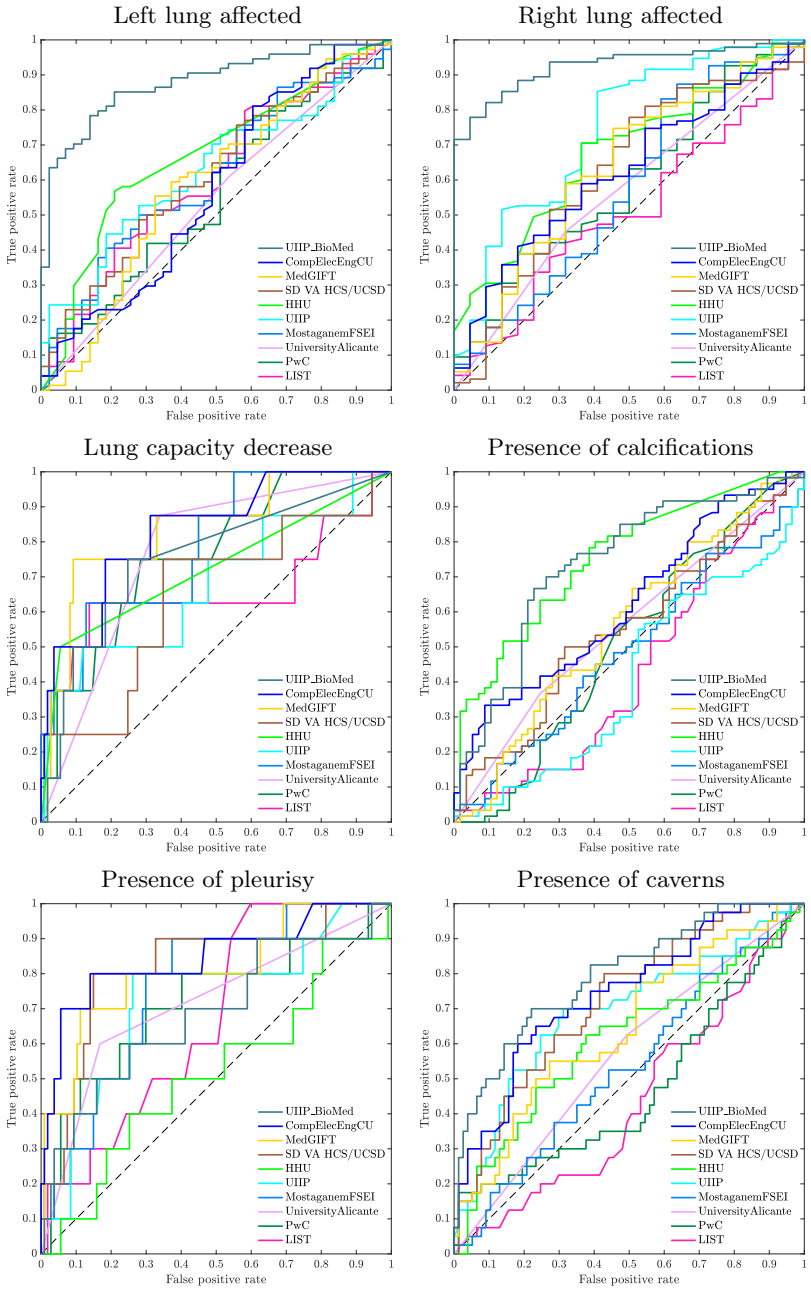

Fig. 3. Receiver operating characteristic (ROC) curves obtained by the best run of each group for each CT finding. The dashed line marks the curve of a random classifier.

Table 6. Detailed results obtained in the CTR task by the best run of each group.

	AUC			Sensitivity			Specificity		
Group name	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
UIIP_BioMed	**0.7968**	**0.6860**	**0.9254**	0.5550	0.1250	0.9579	**0.7398**	0.3636	0.9817
CompElecEngCU	0.7066	0.5739	0.8467	0.5719	0.2000	**1.0000**	0.5948	0.0000	0.9720
MedGIFT	0.6795	0.5626	0.8360	0.3375	0.0000	**1.0000**	0.6667	0.0000	**1.0000**
SD VA HCS/UCSD	0.6631	0.5541	0.8206	0.4936	0.2000	0.9474	0.6301	0.0000	0.9908
HHU	0.6591	0.5159	0.7554	0.4931	0.0000	**1.0000**	0.6452	0.0000	**1.0000**
UIIP	0.6464	0.4099	0.7440	0.4955	0.0000	**1.0000**	0.5155	0.0000	**1.0000**
MostaganemFSEI	0.6273	0.4877	0.7856	0.5109	0.0333	0.9189	0.6394	0.0698	0.9825
UniversityAlicante	0.6190	0.5366	0.7678	**0.5879**	**0.3667**	0.8750	0.6500	**0.4651**	0.8318
PwC	0.6002	0.4724	0.7597	0.4157	0.0000	**1.0000**	0.6457	0.0000	**1.0000**
LIST	0.5523	0.4317	0.6738	0.3816	0.0000	0.9684	0.6760	0.0455	**1.0000**

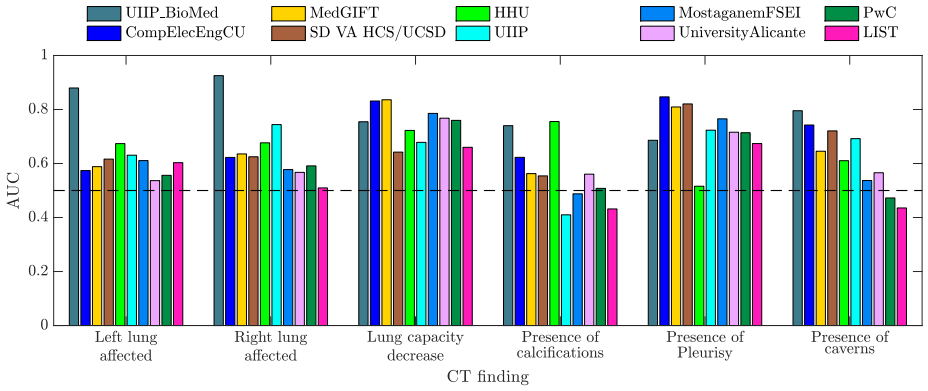

Fig. 4. Area under the ROC curve (AUC) obtained by the best run of each group for each CT finding. The dashed line marks the AUC of a random classifier, 0.50.

The results obtained in this first edition of the CTR subtask showed impressive performance by the participants. Already combining only the best runs analyzed in this work, the AUC for each CT-finding would be of 0.8796, 0.9254, 0.8360, 0.7554, 0.8467 and 0.7955, respectively. However, the sensitivity and specificity seem to not correlate with the AUCs obtained. This is due to the lack of optimization of the classification decision threshold (fixed to 0.50 in our analysis) and this also explains the inverse behavior between specificity and sensitivity of most of the methods. Since the ranking in the CTR task was announced to be evaluated only by AUC, adjusting the decision threshold was not required and hence we assume that no participant adapted the predictions. At the same time, this suggests that maybe AUC was not the best metric to evaluate/rank the methods. A priori, it seems that the AUC only provided information about whether the methods of the participants were capable of ordering the predictions, *i.e.* that a patient with an abnormality presents higher positive probability than a patient without it, but this does not assure that the method is able to distin-

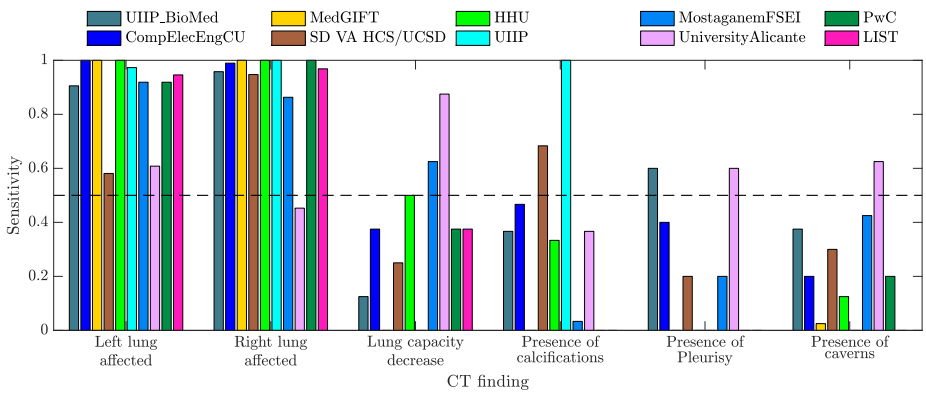

Fig. 5. Sensitivity (true positive rate) obtained by the best run of each group for each CT finding. The dashed line marks the sensitivity of a random classifier, 0.50.

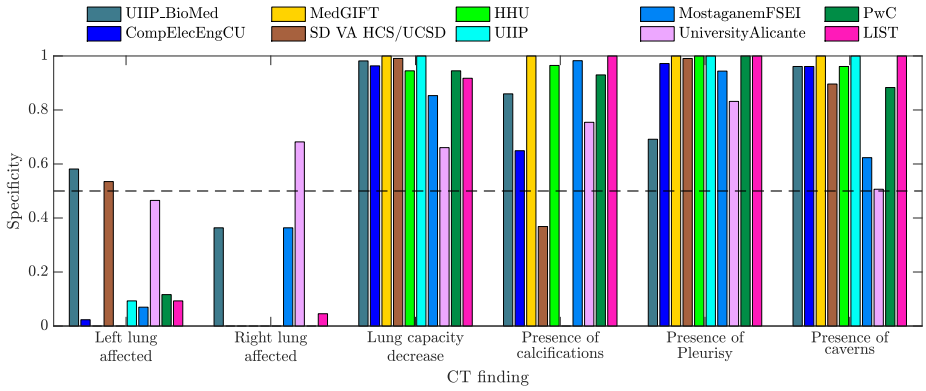

Fig. 6. Specificity (true negative rate) obtained by the best run of each group for each CT finding. The dashed line marks the specificity of a random classifier, 0.50.

guish between presence and absence of a certain CT-finding. Something worth mentioning is the misalignment between the training and test sets in terms of the proportion of positive patients in some of the CT-findings, *e.g.* lung capacity decrease (29-7%) and presence of calcifications (13-51%) (see Table 1). Preserving the proportions for all the CT-findings simultaneously was a extremely difficult task due to the relative small size of the data set. We believe that this misalignment interfered with the generalization power of some methods.

The participants developed many different approaches in 2019, with many of them applying deep learning (DL) techniques. This is actually representative of the current trends in the medical imaging community where DL methods are gaining terrain in almost every area. However, some of the preliminary analysis performed on the CT images by the participants proved that it is more important to understand the problem than to have powerful methods. These analysis led

to simple approaches for some of the abnormalities that resulted in high performance (*e.g.* assessing the presence of pleurisity by comparing lung segmentation masks). We also noticed that all participants model the CTR subtask as a multi-binary problem, with few groups adding a relation between the abnormalities. This was expected since the dataset was not large enough to model the CTR subtask as a multi-label problem due to the high variability when having six labels to predict. Nonetheless, we find surprising that only few groups used their predictions of the CT-findings in their assessment of the TB severity score.

Overall, the 2019 edition of the ImageCLEF TB task again proved the high interest by the medical imaging community in this task resulting in the highest participation of the three editions. Moreover, the results once again support the benefits of applying machine learning techniques in the assessment of a TB severity score, and more precisely in the detection of TB-associated abnormalities. The use of a unique data set for the two tasks allowed to provide a rich set of meta-data for all the patients that was used by most of the participants. However, providing such meta-data affected the size of the data set. In future editions of this task we will focus on extending the data set without reducing the amount of meta-data provided.

Acknowledgements

This work was partly supported by the Swiss National Science Foundation in the project PH4D (320030–146804) and by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project DAA3-18-64818-1 "Year 7: Belarus TB Database and TB Portals".

References

1. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0) (2015) 55 – 61
2. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF – Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Springer International Series On Information Retrieval. Springer, Berlin Heidelberg (2010)
3. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum). (September 2016)
4. Müller, H., Clough, P., Hersh, W., Geissbuhler, A.: ImageCLEF 2004–2005: Results experiences and new ideas for image retrieval evaluation. In: International Conference on Content–Based Multimedia Indexing (CBMI 2005), Riga, Latvia, IEEE (June 2005)
5. Ionescu, B., Müller, H., Péteri, R., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M.,

Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Volume 2380 of Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019)., Lugano, Switzerland, LNCS Lecture Notes in Computer Science, Springer (September 9-12 2019)

6. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France, LNCS Lecture Notes in Computer Science, Springer (September 10-14 2018)
7. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Volume 10456 of Lecture Notes in Computer Science., Dublin, Ireland, Springer (September 11-14 2017)
8. Villegas, M., Müller, H., Garcia Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, A., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sanchez, J.A., Vidal, E.: General overview of ImageCLEF at the CLEF 2016 labs. In: CLEF 2016 Proceedings. Lecture Notes in Computer Science, Evora. Portugal, Springer (September 2016)
9. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science. Springer International Publishing (2015)
10. Caputo, B., Müller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Martinez Gomez, J., Garcia Varea, I., Cazorla, C.: ImageCLEF 2013: the vision, the data and the open challenges. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum). (September 2013)
11. World Health Organization, et al.: Global tuberculosis report 2016. (2016)
12. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org <http://ceur-ws.org> (September 11-14 2017)
13. Dicente Cid, Y., Jimenez-del-Toro, O., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in CT volumes. In Orcun Goksel, Jimenez-del-Toro, O., Foncubierta-Rodriguez, A., Müller, H., eds.: Proceedings of the VISCERAL Challenge at ISBI. Number 1390 in CEUR Workshop Proceedings (Apr 2015) 31–35
14. Bendayan, D., Barziv, Y., Kramer, M.: Pulmonary calcifications: a review. Respiratory medicine **94**(3) (2000) 190–3

15. Gadkowski, L.B., Stout, J.E.: Cavitary pulmonary disease. Clinical Microbiology Reviews **21**(2) (2008) 305–333
16. Liauchuk, V.: ImageCLEF 2019: Projection-based CT Image Analysis for TB Severity Scoring and CT Report Generation. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
17. Kazlouski, S.: ImageCLEF 2019: CT Image Analysis for TB Severity Scoring and CT Report Generation using Autoencoded Image Features. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
18. Bogomasov, K., Braun, D., Burbach, A., Himmelspach, L., Conrad, S.: Feature and Deep Learning Based Approaches for Automatic Report Generation and Severity Scoring of Lung Tuberculosis from CT Images. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
19. Gentili, A.: ImageCLEF2019: Tuberculosis - Severity Scoring and CT Report with Neural Networks, Transfer Learning and Ensembling. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
20. Dicente Cid, Y., Müller, H.: Lung Graph-Model Classification with SVM and CNN for Tuberculosis Severity Assessment and Automatic CT Report Generation. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
21. Hamadi, A., Cheikh, N.B., Zouatine, Y., Menad, S.M.B.: ImageCLEF 2019: Deep Learning for Tuberculosis CT Image Analysis. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
22. Mossa, A.A., Yibre, A.M., Çevik, U.: Multi-View CNN with MLP for Diagnosing Tuberculosis Patients Using CT Scans and Clinically Relevant Metadata. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
23. Llopis, F., Fuster, A., Azorín, J., Llopis, I.: Using improved optical flow model to detect Tuberculosis. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
24. Kavitha, S., Nandhinee, P., Harshana, S., Jahnavi Srividya, S., Harrinei, K.: ImageCLEF 2019: A 2D Convolutional Neural Network Approach for Severity Scoring of Lung Tuberculosis using CT Images. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
25. Zunair, H., Rahman, A., Mohammed, N.: Estimating Severity from CT Scans of Tuberculosis Patients using 3D Convolutional Nets and Slice Selection. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
26. Tabarcea, A., Rosca, V., Iftene, A.: ImageCLEFmed Tuberculosis 2019: Predicting CT Scans Severity Scores using Stage-Wise Boosting in Low-Resource Environments. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Pro-

ceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)
27. Pattnaik, A., Kanodia, S., Chowdhury, R., Mohanty, S.: Predicting Tuberculosis Related Lung Deformities from CT Scan Images Using 3D CNN. In: CLEF2019 Working Notes. Volume 2380 of CEUR Workshop Proceedings., Lugano, Switzerland, CEUR-WS.org <http://ceur-ws.org/Vol-2380> (September 9-12 2019)