Published in "Proceedings of CLEF 2018 Working Notes", 2018, 10-14 September, Avignon, France, which should be cited to refer to this work.

Overview of ImageCLEFtuberculosis 2018 – Detecting Multi-Drug Resistance, Classifying Tuberculosis Types and Assessing Severity Scores

Yashin Dicente Cid[1,2], Vitali Liauchuk[3], Vassili Kovalev[3], and Henning Müller[1,2]

[1] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
[2] University of Geneva, Switzerland;
[3] United Institute of Informatics Problems, Minsk, Belarus
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. The tuberculosis task was held for the first time in 2017 and had a very encouraging participation with 9 groups submitting results to these very challenging tasks. In 2018 there was a slightly higher participation. Three tasks were proposed in 2018: (1) the detection of drug resistances among tuberculosis cases, (2) the classification of the cases into five types of tuberculosis and (3) the assessment of a tuberculosis severity score. Many different techniques were used by the participants ranging from Deep Learning to graph-based approaches and best results were obtained by a variety of approaches with no clear technique dominating. Both, the detection of drug resistances and the classification of tuberculosis types had similar results than in the previous edition, the former remaining as a very difficult task. In the case of the severity score task, the results support the suitability of assessing the severity based only on the CT image, as the results obtained were very good.

Keywords: Tuberculosis, Computed Tomography, Image Classification, Drug Resistance, Severity Scoring, 3D Data Analysis

1 Introduction

ImageCLEF[4] 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

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

on the other tasks organized in 2018 can be found in [5] and the past editions are described in [6–9].

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 [10]. This bacteria usually attacks the lungs, but it can also damage other parts of the body. Generally, TB can be cured with antibiotics. However, the greatest disaster 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 [11], 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 datasets. Moreover, a new subtask was added based on assessing a severity score of the disease given a CT image.

This article first describes the three tasks proposed around TB in 2018. Then, the datasets, evaluation methodology and participation are detailed. The results section describes the submitted runs and the results obtained for the three subtasks. A discussion and conclusion section ends the paper.

2 Tasks, Datasets, Evaluation, Participation

2.1 The Tasks in 2018

Three subtasks were organized in 2018. Two were common with the 2017 edition and one new subtask was added:

- Multi-Drug Resistance detection (MDR subtask);
- Tuberculosis Type classification (TBT subtask);
- Severity Scoring assessment (SVR subtask).

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

Multi-drug Resistance Detection: As in 2017, the goal of the MDR subtask was to assess the probability of a TB patient having a resistant form of TB based on the analysis of a chest CT scan alone. The dataset for this subtask was increased from the previous year but the subtask remained as a binary classification problem even though several levels of resistances exist.

Tuberculosis Type Classification: This subtask is also common with the 2017 edition and, like in the MDR subtask, we increased the dataset. The goal of the TBT subtask is to automatically categorize each TB case into one of the following five TB types: Infiltrative, Focal, Tuberculoma, Miliary, and Fibrocavernous. The distribution of cases among the classes is not balanced but the distributions are similar in the training and the test data.

Severity Scoring: This subtask aims at assessing a TB severity score based only on a chest CT image. 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"). In the process of scoring, the medical doctors considered many factors like pattern of the lesions, results of microbiological tests, duration of treatment, patient age and other criteria.

2.2 Datasets

For each of the three subtasks, a separate dataset was provided, all containing 3D CT images stored in the NIfTI (Neuroimaging Informatics Technology Initiative) file format with slice resolution of 512×512 pixels and a number of slices varying from about 50 to 400. A set of relevant meta-data such as age and gender was provided for each subtask. The entire dataset including CT images and 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 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[5] 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.

In the framework of the ImageCLEF 2018 TB task, automatically extracted masks of the lungs were provided for all CT images. These masks were extracted using the method described in [12]. The segmentations were analyzed based on the number of lungs found and the size ratio of the lungs in a supervised manner. Only those segmentations with anomalies on these two metrics were visualized and evaluated accordingly. A total of 32 images out of 2,287 presented a problematic mask, 8 including areas outside the lungs and 24 containing only one lung. The 8 inaccurate masks were corrected by fusing the above mentioned method and the registration-based segmentation used in [13]. The other 24 masks (20 from the TBT subtask and 4 from the MDR subtask) could not be properly

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

labeled due to the size and/or damage of one lung. In these cases, the masks provided to the participants only contained one label (right lung).

Multi-drug Resistance Detection The dataset for this task is an extension of the one used in the 2017 edition. Particularly, the training and test sets of this subtask were extended by adding patients with extensively drug-resistant (XDR) TB, which is a rare and more severe subtype of MDR TB. Along with the 3D CT images and lung masks, the age and gender of each patient were provided. The dataset includes only HIV-negative patients with no relapses. Each patient was classified into one the two classes: drug sensitive (DS) or multi-drug resistant (MDR). A patient was considered DS if the TB bacteria was sensitive to all the anti-tuberculosis drugs tested. All XDR patients were considered to belong to the MDR class. Table 1 contains the number of patients in each set.

Table 1. Dataset of the MDR detection subtask.

Patient set	**Train**	**Test**
DS	134	99
MDR	125	137
Total patients	**259**	**236**

Tuberculosis Type Classification The dataset used in this subtask includes chest CT scans of TB patients along with the TB type and patient age at the moment of the scan. Like the MDR dataset, the TBT 2017 dataset was extended for the 2018 edition. In this case, new CT scans of the same patients involved in 2017 were added and also some CT images of new patients. In the TBT 2018 dataset, for each patient there are between 1 and 9 CT scans acquired at different time points. All scans of the same patient were diagnosed with the same TB type by expert radiologists. Figure 1 shows one example for each of the five TB types. Moreover, Figure 2 shows examples of two patients with three CT scans each. The CT slices in both figures are shown using a Hounsfield Unit (HU) window with center at -500 HU and width of 1400 HU. The number of CT scans and patients in each TB type set are shown in Table 2.

Severity Scoring The data for the SVR subtask includes 279 CT scans with known TB severity scores ranging from 1 to 5 assigned by medical doctors. Each CT scan corresponds to a specific TB patient. To treat this subtask as a binary classification problem, the severity scores were grouped so that values 1, 2 and 3 corresponded to "high severity" class, and values 4 and 5 corresponded to "low severity". Table 3 contains the number of patients of each severity class in the sets.

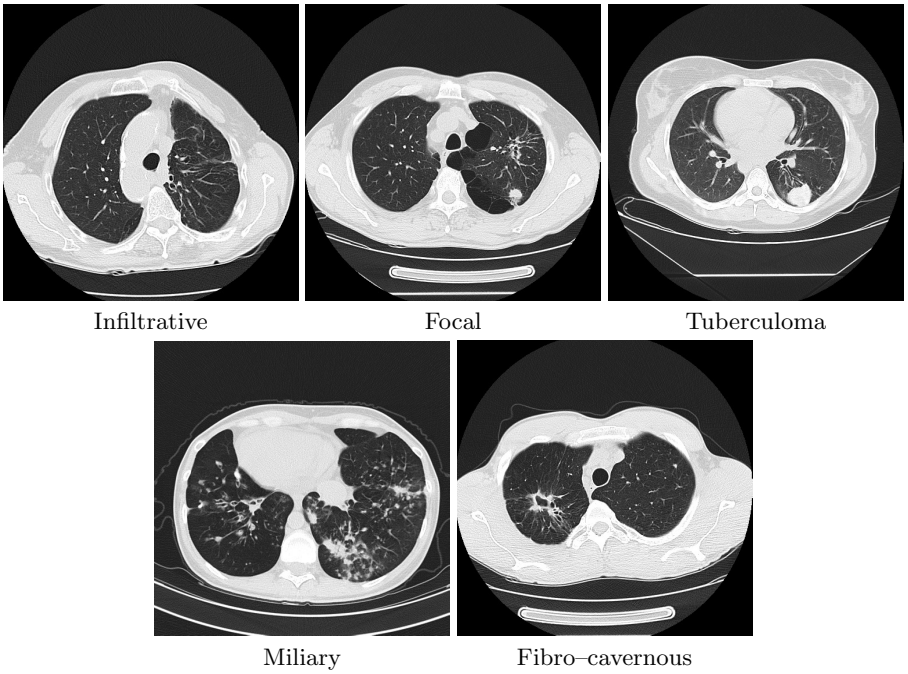

Fig. 1. Examples of the five TB types in the TBT subtask. The CT slices are shown using a HU window with center at -500 HU and width of 1400 HU.

2.3 Evaluation Measures and Scenario

Similar to 2017, the participants were allowed to submit up to 10 runs to each of the three TB subtasks. In the case of the MDR subtask, the participants had to

Table 2. Dataset of the TBT classification subtask.

	Num. Patients (CT series)			
Patient set	**Train**		**Test**	
Type 1 (T1) – Infiltrative	228	(376)	89	(179)
Type 2 (T2) – Focal	210	(273)	80	(115)
Type 3 (T3) – Tuberculoma	100	(154)	60	(86)
Type 4 (T4) – Miliary	79	(106)	50	(71)
Type 5 (T5) – Fibro-cavernous	60	(99)	38	(57)
Total patients (CTs)	**677**	**(1,008)**	**317**	**(505)**

Table 3. Dataset of the SVR subtask.

Patient set	**Train**	**Test**
Low severity	90	62
High severity	80	47
Total patients	**170**	**109**

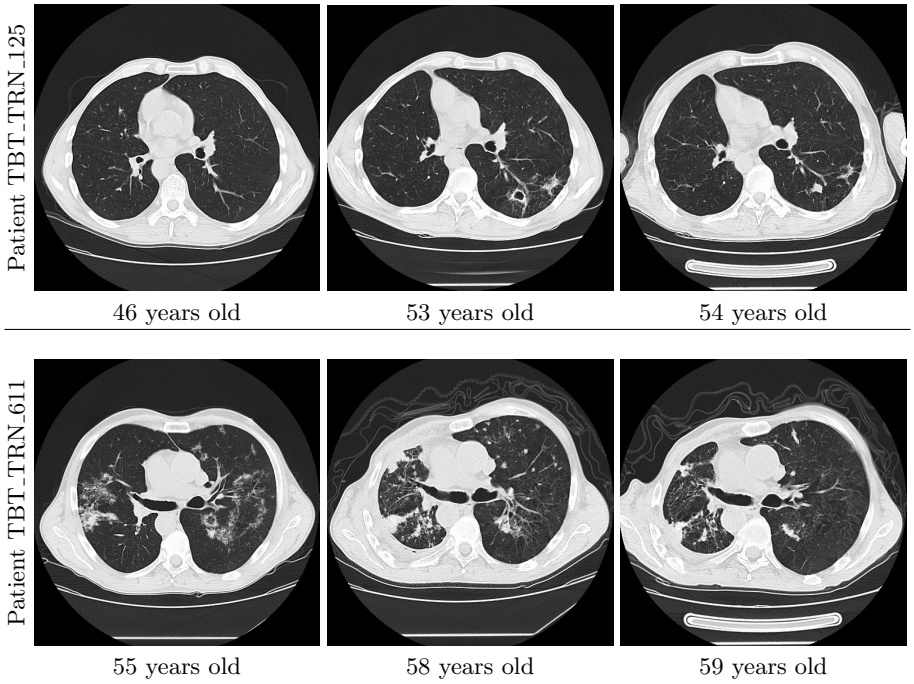

Fig. 2. Examples of two patients (TBT_TRN_125 and TBT_TRN_611) in the TBT dataset with three CT scans taken at different points in time. Each row contains a slice of the three scans of a patient ordered by the time it was taken. The three CT images of patient TBT_TRN_125 were classified as having TB type 1 (infiltrative) while the three series of patient TBT_TRN_611 are of type 4 (miliary). All images are shown using a HU window with center at -500 HU and width of 1400 HU.

provide the probability for the TB cases to belong to the MDR class ranging from 0 to 1. These probabilities were used to build Receiver Operating Characteristic (ROC) curves. Since the MDR dataset was not perfectly balanced and had a relatively small size, Area Under the ROC Curve (AUC) was used to evaluate the participant runs. We provided the accuracy of the binary classification using a standard threshold of 0.50.

In the case of the TBT task, the participants had to predict the TB type of each patient, and submit a run containing a category label in the set {1, 2, 3, 4, 5}. Considering that a high number of patients in the dataset had multiple CT scans with the same TB type, the evaluation was performed patient-wise. Cohen's Kappa coefficient was provided for each run along with the 5-class prediction accuracy. Cohen's Kappa is not sensitive to unbalanced datasets, which is the case for the data used here.

The runs submitted for the severity scoring subtask were evaluated in two ways. One used the original severity scores from 1 to 5 and the task was to predict those numerical scores as precise as possible (a regression problem).

Here, Root Mean Square Error (RMSE) was computed between the ground truth severity and the predicted scores provided by the participants. Alternatively, the original severity score was transformed into two classes, where scores from 1 to 3 corresponded to "high severity" and the 4 and 5 scores corresponded to the "low severity" class. In this case the participants had to provide the probability of TB cases to belong to the "high severity" class. The corresponding results were evaluated using AUC.

2.4 Participation

In 2018 there were 85 registered teams and 33 signed the end user agreement. Finally, 11 groups from 9 countries participated in one or more subtasks and submitted results. These numbers are similar to 2017, where there were 94 registered teams, 48 that signed the end user agreement, and 9 teams from 9 countries submitting results. Table 4 shows the list of participants and the subtasks where they participated. One of the groups (HHU-DBS) participated in two subtasks with different approaches developed by a different set of authors. Therefore, their approaches are referred as HHU-DBS_1 and HHU-DBS_2 in the following sections.

Table 4. List of participants submitting a run to at least one subtask.

			Subtask		
Group name	**Main institution**	**Country**	**MDR**	**TBT**	**SVR**
fau_ml4cv	Florida Atlantic University	USA	–	×	–
HHU-DBS (*)	Heinrich Heine University	Germany	×	–	×
LIST	Abdelmalek Essadi University	Morocco	×	×	–
MedGIFT	University of Applied Sciences Western Switzerland (HES–SO)	Switzerland	×	×	×
Middlesex University	Middlesex University	UK	–	–	×
MostaganemFSEI	University of Abdelhamid Ibn Badis Mostaganem	Algeria	–	×	×
SD VA HCS/UCSD	San Diego VA Health Care System	USA	×	×	×
UIIP_BioMed	United Institute of Informatics Problems	Belarus	×	×	×
UniversityAlicante	University of Alicante	Spain	×	×	–
VISTA@UEvora	University of Évora	Portugal	×	×	×

(*) The HHU-DBS group participated with different approaches in the MDR and SVR subtasks. Therefore, the group name is split into HHU-DBS_1 and HHU-DBS_2 respectively in the following sections.

3 Results

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

3.1 MDR Detection

Table 5 shows the results obtained for the MDR detection subtask. The runs were evaluated using ROC curves produced from the probabilities provided by the participants. The results in the table are sorted by AUC in descending order. The accuracy is given in the table as well. Additionally, Figure 3 shows the highest AUC values achieved by the participants compared to the best result obtained in the 2017 edition.

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

Group name	Run	AUC	Rank AUC	Acc	Rank Acc
VISTA@UEvora	MDR-Run-06-Mohan-SL-F3-Personal.txt	**0.6178**	1	0.5593	8
SD VA HCS/UCSD	MDSTest1a.csv	0.6114	2	**0.6144**	1
VISTA@UEvora	MDR-Run-08-Mohan-voteLdaSmoF7-Personal.txt	0.6065	3	0.5424	17
VISTA@UEvora	MDR-Run-09-Sk-SL-F10-Personal.txt	0.5921	4	0.5763	3
VISTA@UEvora	MDR-Run-10-Mix-voteLdaSl-F7-Personal.txt	0.5824	5	0.5593	9
HHU-DBS_1	MDR_FlattenCNN_DTree.txt	0.5810	6	0.5720	4
HHU-DBS_1	MDR_FlattenCNN2_DTree.txt	0.5810	7	0.5720	5
HHU-DBS_1	MDR_Conv68adam_fl.txt	0.5768	8	0.5593	10
VISTA@UEvora	MDR-Run-07-Sk-LDA-F7-Personal.txt	0.5730	9	0.5424	18
UniversityAlicante	MDRBaseline0.csv	0.5669	10	0.4873	32
HHU-DBS_1	MDR_Conv48sgd.txt	0.5640	11	0.5466	16
HHU-DBS_1	MDR_Flatten.txt	0.5637	12	0.5678	7
HHU-DBS_1	MDR_Flatten3.txt	0.5575	13	0.5593	11
UIIP_BioMed	MDR_run_TBdescs2_zparts3_thrprob50_rf150.csv	0.5558	14	0.4576	36
UniversityAlicante	testSVM_SMOTE.csv	0.5509	15	0.5339	20
UniversityAlicante	testOpticalFlowwFrequencyNormalized.csv	0.5473	16	0.5127	24
HHU-DBS_1	MDR_Conv48sgd_fl.txt	0.5424	17	0.5508	15
HHU-DBS_1	MDR_CustomCNN_DTree.txt	0.5346	18	0.5085	26
HHU-DBS_1	MDR_FlattenX.txt	0.5322	19	0.5127	25
HHU-DBS_1	MDR_MultiInputCNN.txt	0.5274	20	0.5551	13
VISTA@UEvora	MDR-Run-01-sk-LDA.txt	0.5260	21	0.5042	28
MedGIFT	MDR_Riesz_std_correlation_TST.csv	0.5237	22	0.5593	12
MedGIFT	MDR_HOG_std_euclidean_TST.csv	0.5205	23	0.5932	2
VISTA@UEvora	MDR-Run-05-Mohan-RF-F3I650.txt	0.5116	24	0.4958	30
MedGIFT	MDR_AllFeats_std_correlation_TST.csv	0.5095	25	0.4873	33
UniversityAlicante	DecisionTree25v2.csv	0.5049	26	0.5000	29
MedGIFT	MDR_AllFeats_std_euclidean_TST.csv	0.5039	27	0.5424	19
LIST	MDRLIST.txt	0.5029	28	0.4576	37
UniversityAlicante	testOFFullVersion2.csv	0.4971	29	0.4958	31
MedGIFT	MDR_HOG_mean_correlation_TST.csv	0.4941	30	0.5551	14
MedGIFT	MDR_Riesz_AllCols_correlation_TST.csv	0.4855	31	0.5212	22
UniversityAlicante	testOpticalFlowFull.csv	0.4845	32	0.5169	23
MedGIFT	MDR_Riesz_mean_euclidean_TST.csv	0.4824	33	0.5297	21
UniversityAlicante	testFrequency.csv	0.4781	34	0.4788	34
UniversityAlicante	testflowI.csv	0.4740	35	0.4492	39
MedGIFT	MDR_HOG_AllCols_euclidean_TST.csv	0.4693	36	0.5720	6
VISTA@UEvora	MDR-Run-06-Sk-SL.txt	0.4661	37	0.4619	35
MedGIFT	MDR_AllFeats_AllCols_correlation_TST.csv	0.4568	38	0.5085	27
VISTA@UEvora	MDR-Run-04-Mix-Vote-L-RT-RF.txt	0.4494	39	0.4576	38

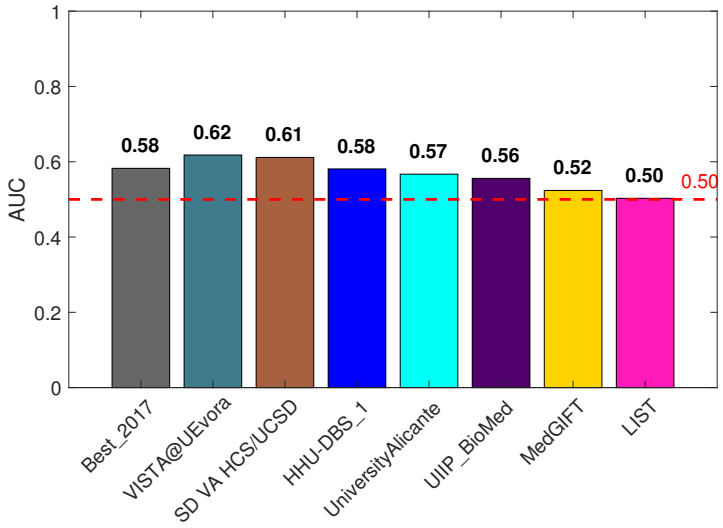

Fig. 3. Area Under the ROC Curve (AUC) obtained by the best run of each group. "Best_2017" corresponds to the best AUC obtained in the 2017 edition. The red line marks the baseline of 0.50 AUC corresponding to a random classifier.

It is worth to notice that the image-based detection task of MDR TB remains very challenging and so far has no solution with a sufficiently high prediction accuracy for being useful in clinical practice. Recent articles report the presence of statistically significant links between drug resistance and multiple thick-walled caverns [14]. However, computerized methods show a performance of image-based MDR TB detection barely beyond the level of statistical significance compared to a random classifier [6, 15, 16].

The best result in terms of AUC was achieved by VISTA@UEvora team with an AUC of 0.6178 [17]. The team used conventional approaches for the extraction of quantitative image descriptors, such as statistical moments, fractal dimension, gray-level co-occurrence matrices and their derivative features. A set of conventional classification methods was used for prediction in all the three subtasks. Their best run in terms of classification accuracy (0.5763) ranked 3rd place among the participant runs and is not the same run that had the best AUC. The second highest AUC of 0.6114 was achieved by the San Diego VA HCS/UCSD [18] with an approach based on splitting the 3D CT scans into a set of 2D images and using a pre-trained ResNeXt deep network for classification. This run achieved the highest MDR detection accuracy (0.6144). The third highest AUC was obtained by HHU-DBS_1 [19]. They used 3D deep Convolutional Neural Networks (CNNs) combined with decision trees and obtained 0.5810 AUC and 0.5720 classification accuracy with their best run. The UniversityAlicante group used two approaches: one based on 2D CNNs and the other based on Optical Flow (OF) [20]. The best AUC among this group's runs was obtained using

only patient age and gender information and ranked 10th among all participant runs with a AUC of 0.5669. Other runs obtained lower AUC. This OF-based approach for CT image analysis resulted in an accuracy of 0.5339 and ranked 20th. The single run submitted by the UIIP_BioMed group ranked 14th in AUC and 36th in accuracy with an AUC of 0.5558 and an accuracy of 0.4576 [21]. A technique for automatic detection of lesions of different types in a six-region division of the CT lung volume was used. A separate dataset with labeled lesions in CT was used for training the lesion detection algorithm. A Random Forest (RF) classifier was used for the prediction of the final classes and scores in all three subtasks. Methods based on a graph-model of the lungs and 3D texture analysis were used by MedGIFT group [22]. Their best runs resulted in the 22nd highest AUC (0.5237) and the 2nd highest accuracy (0.5932). Finally, the LIST group used a hybrid approach that combined 3D CNNs with linear SVM classifiers for MDR detection and TB type classification [23]. The single run submitted by the group obtained an AUC of 0.5029 and an accuracy of 0.4576 and ranked 28th and 37th, respectively. The information about age and gender of TB patients was used only by two participating groups: HHU-DBS_1 and UIIP_BioMed.

3.2 Tuberculosis Type Classification

Table 6 shows the results obtained for the TBT subtask. The runs were evaluated on the test set of images using the unweighted Cohen Kappa coefficient and overall classification accuracy. The results are sorted by Cohen's Kappa in descending order. Figure 4 shows the highest Kappa values achieved by the participants. The true positive rates of the different TB types are shown in Figure 5.

In the TBT subtask, most of the teams used the same methods as they used for the MDR detection. The best result in terms of both Kappa and classification accuracy was achieved by the UIIP_BioMed group with the use of a lesion-based TB descriptor and a RF classifier. The run resulted in a Kappa of 0.2312 and a classification accuracy of 0.4227. Instead of using all the available CT series, this group only used the first scan of a patient for the classification of the TB type. The second highest Kappa was achieved by the fau_ml4cv group that participated only in the TBT subtask [24]. An ensemble of 3D CNNs was used, achieving a Kappa of 0.1736 and an accuracy of 0.3533 with their best run. The graph-based approach of the MedGIFT team resulted in the 2nd best classification accuracy (0.3849) and the 3rd highest Kappa (0.1706). The best runs of VISTA@UEvora, San Diego VA HCS/UCSD, UniversityAlicante and LIST resulted in Kappa values of 0.1664, 0.1474, 0.0204, and -0.0024 respectively. The MostaganemFSEI group participated in the TBT classification and the SVR subtasks. The algorithm employed by them was based on splitting the 3D CT scans into 2D slices, extracting semantic descriptors using a trained CNN and applying conventional classification methods [25]. They obtained a Kappa of 0.0629 and an accuracy of 0.2744. It is worth to highlight that only the fau_ml4cv and San Diego VA HCS/UCSD groups obtained a true positive rate higher than a random classifier in all five TB types (see Figure 5).

3.3 Severity Score

The results obtained for the severity scoring subtask are shown in Table 7. The best RMSE achieved by the participating groups and the corresponding AUCs are shown in Figures 6 and 7. The best results in terms of regression were obtained by the UIIP_BioMed group with an RMSE of 0.7840, which also achieved the 6th best classification result with an AUC of 0.7025. The highest classification result was achieved by the MedGIFT group with an AUC of 0.7708. The MedGIFT group's best regression obtained an RMSE of 0.8513, which is the second best result. The third best RMSE (0.8883) was obtained by the VISTA@UEvora group. The same run ranked on the 21st place for classification

Table 6. Results obtained by the participants in the TBT task.

Group name	Run	Kappa	Rank Kappa	Acc	Rank Acc
UIIP_BioMed	TBT_run_TBdescs2_zparts3_thrprob50_rf150.csv	**0.2312**	1	**0.4227**	1
fau_ml4cv	TBT_m4_weighted.txt	0.1736	2	0.3533	10
MedGIFT	TBT_AllFeats_std_euclidean_TST.csv	0.1706	3	0.3849	2
MedGIFT	TBT_Riesz_AllCols_euclidean_TST.csv	0.1674	4	0.3849	3
VISTA@UEvora	TBT-Run-02-Mohan-RF-F20I1500S20-317.txt	0.1664	5	0.3785	4
fau_ml4cv	TBT_m3_weighted.txt	0.1655	6	0.3438	12
VISTA@UEvora	TBT-Run-05-Mohan-RF-F20I2000S20.txt	0.1621	7	0.3754	5
MedGIFT	TBT_AllFeats_AllCols_correlation_TST.csv	0.1531	8	0.3691	7
MedGIFT	TBT_AllFeats_mean_euclidean_TST.csv	0.1517	9	0.3628	8
MedGIFT	TBT_Riesz_std_euclidean_TST.csv	0.1494	10	0.3722	6
SD VA HCS/UCSD	Task2Submission64a.csv	0.1474	11	0.3375	13
SD VA HCS/UCSD	TBTTask_2_128.csv	0.1454	12	0.3312	15
MedGIFT	TBT_AllFeats_AllCols_correlation_TST.csv	0.1356	13	0.3628	9
VISTA@UEvora	TBT-Run-03-Mohan-RF-7FF20I1500S20-Age.txt	0.1335	14	0.3502	11
SD VA HCS/UCSD	TBTLast.csv	0.1251	15	0.3155	20
fau_ml4cv	TBT_w_combined.txt	0.1112	16	0.3028	22
VISTA@UEvora	TBT-Run-06-Mix-RF-5FF20I2000S20.txt	0.1005	17	0.3312	16
VISTA@UEvora	TBT-Run-04-Mohan-VoteRFLMT-7F.txt	0.0998	18	0.3186	19
MedGIFT	TBT_HOG_AllCols_euclidean_TST.csv	0.0949	19	0.3344	14
fau_ml4cv	TBT_combined.txt	0.0898	20	0.2997	23
MedGIFT	TBT_HOG_std_correlation_TST.csv	0.0855	21	0.3218	18
fau_ml4cv	TBT_m2p01_small.txt	0.0839	22	0.2965	25
MedGIFT	TBT_AllFeats_std_correlation_TST.csv	0.0787	23	0.3281	17
fau_ml4cv	TBT_m2.txt	0.0749	24	0.2997	24
MostaganemFSEI	TBT_mostaganemFSEI_run4.txt	0.0629	25	0.2744	27
MedGIFT	TBT_HOG_std_correlation_TST.csv	0.0589	26	0.3060	21
fau_ml4cv	TBT_modelsimple_lmbdap1_norm.txt	0.0504	27	0.2839	26
MostaganemFSEI	TBT_mostaganemFSEI_run1.txt	0.0412	28	0.2650	29
MostaganemFSEI	TBT_MostaganemFSEI_run2.txt	0.0275	29	0.2555	32
MostaganemFSEI	TBT_MostaganemFSEI_run6.txt	0.0210	30	0.2429	33
UniversityAlicante	3nnconProbabilidad2.txt	0.0204	31	0.2587	30
UniversityAlicante	T23nnFinal.txt	0.0204	32	0.2587	31
fau_ml4cv	TBT_m1.txt	0.0202	33	0.2713	28
LIST	TBTLIST.txt	-0.0024	34	0.2366	34
MostaganemFSEI	TBT_mostaganemFSEI_run3.txt	-0.0260	35	0.1514	37
VISTA@UEvora	TBT-Run-01-sk-LDA-Update-317-New.txt	-0.0398	36	0.2240	35
VISTA@UEvora	TBT-Run-01-sk-LDA-Update-317.txt	-0.0634	37	0.1956	36
UniversityAlicante	T2SVMFinal.txt	-0.0920	38	0.1167	38
UniversityAlicante	SVMirene.txt	-0.0923	39	0.1136	39

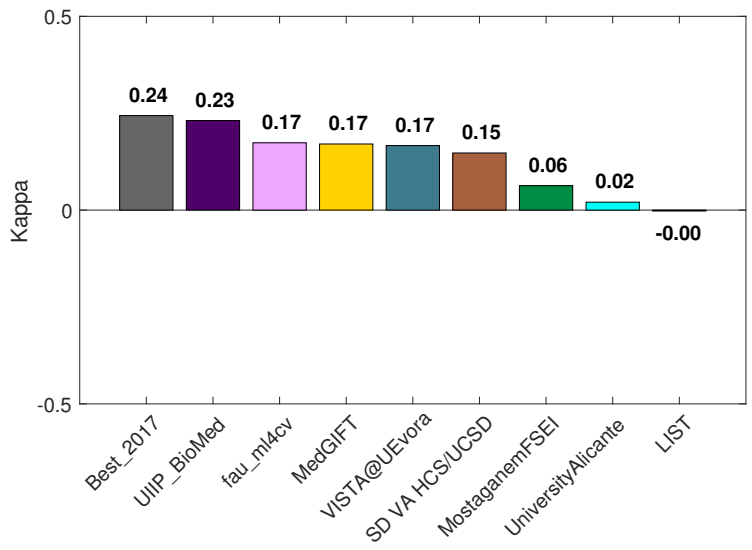

Fig. 4. The unweighted Cohen Kappa coefficient obtained by the best run of each group. "Best_2017" refers to the best run in the 2017 edition.

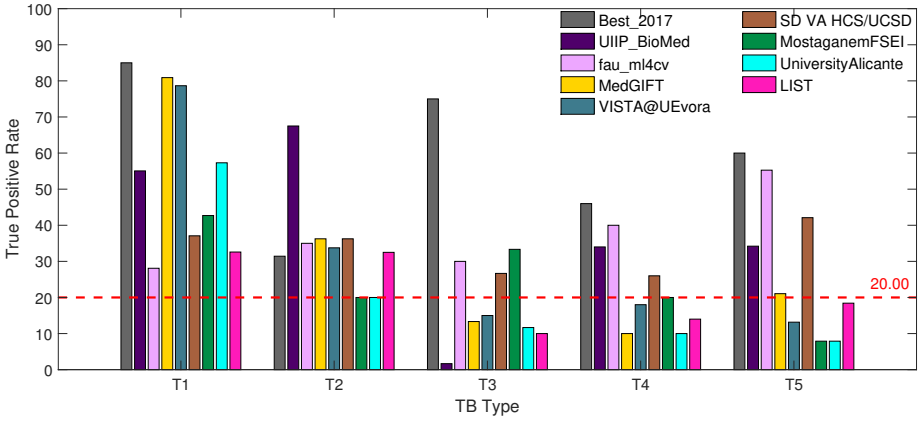

Fig. 5. True positive rate (%) for each TB type obtained by the best run of each group. "Best_2017" refers to the best run in 2017. The red line shows the true positive rate expected for a random classifier in a 5-class problem (20%).

with an AUC of 0.6239. The third best result for classification was obtained by the San Diego VA HCS/UCSD group with an AUC of 0.6984, which corresponds to the 7th best result. Their best regression is an RMSE of 1.2153, which is at rank 30. The HHU-DBS_2 team used a feature-based approach for scoring the severity of TB based on a set of conventional methods [26]. The approach employed image binarization and extraction of features including the presence of

calcifications, lung wateriness, cavities, infection ratio, HU histograms and lung shape to characterize the volumes. The group obtained the 10th best RMSE (0.9626) and 8th best AUC (0.6862). The MostaganemFSEI group achieved an RMSE of 0.9721 and an AUC of 0.6127. Middlesex University participated only in the SVR subtask. The group employed an approach based on using deep residual learning, training on a set of overlapping $128 \times 128 \times$ depth blocks, assessing the TB severity for each block and gathering the results [27]. This allowed to achieve an RMSE of 1.0921 and an AUC of 0.6534 that correspond to the 24th and 14th positions. It is important to highlight that all groups obtained an AUC higher than a random classifier (AUC of 0.50) with all their runs.

Table 7. Results obtained by the participants in the SVR subtask.

Group name	Run	RMSE	Rank RMSE	AUC	Rank AUC
UIIP_BioMed	SVR_run_TBdescs2_zparts3_thrprob50_rf100.csv	**0.7840**	1	0.7025	6
MedGIFT	SVR_HOG_std_euclidean_TST.csv	0.8513	2	0.7162	5
VISTA@UEvora	SVR-Run-07-Mohan-MLP-6FTT100.txt	0.8883	3	0.6239	21
MedGIFT	SVR_AllFeats_AllCols_euclidean_TST.csv	0.8883	4	0.6733	10
MedGIFT	SVR_AllFeats_AllCols_correlation_TST.csv	0.8934	5	**0.7708**	1
MedGIFT	SVR_HOG_mean_euclidean_TST.csv	0.8985	6	0.7443	3
MedGIFT	SVR_HOG_mean_correlation_TST.csv	0.9237	7	0.6450	18
MedGIFT	SVR_HOG_AllCols_euclidean_TST.csv	0.9433	8	0.7268	4
MedGIFT	SVR_HOG_AllCols_correlation_TST.csv	0.9433	9	0.7608	2
HHU-DBS_2	SVR_RanFrst.txt	0.9626	10	0.6484	16
MedGIFT	SVR_Riesz_AllCols_correlation_TST.csv	0.9626	11	0.5535	34
MostaganemFSEI	SVR_mostaganemFSEI_run3.txt	0.9721	12	0.5987	25
HHU-DBS_2	SVR_RanFRST_depth_2_new_new.txt	0.9768	13	0.6620	13
HHU-DBS_2	SVR_LinReg_part.txt	0.9768	14	0.6507	15
MedGIFT	SVR_AllFeats_mean_euclidean_TST.csv	0.9954	15	0.6644	12
MostaganemFSEI	SVR_mostaganemFSEI_run6.txt	1.0046	16	0.6119	23
VISTA@UEvora	SVR-Run-03-Mohan-MLP.txt	1.0091	17	0.6371	19
MostaganemFSEI	SVR_mostaganemFSEI_run4.txt	1.0137	18	0.6107	24
MostaganemFSEI	SVR_mostaganemFSEI_run1.txt	1.0227	19	0.5971	26
MedGIFT	SVR_Riesz_std_correlation_TST.csv	1.0492	20	0.5841	29
VISTA@UEvora	SVR-Run-06-Mohan-VoteMLPSL-5F.txt	1.0536	21	0.6356	20
VISTA@UEvora	SVR-Run-02-Mohan-RF.txt	1.0580	22	0.5813	31
MostaganemFSEI	SVR_mostaganemFSEI_run2.txt	1.0837	23	0.6127	22
Middlesex University	SVR-Gao-May4.txt	1.0921	24	0.6534	14
HHU-DBS_2	SVR_RanFRST_depth_2_Ludmila_new_new.txt	1.1046	25	0.6862	8
VISTA@UEvora	SVR-Run-05-Mohan-RF-3FI300S20.txt	1.1046	26	0.5812	32
VISTA@UEvora	SVR-Run-04-Mohan-RF-F5-I300-S200.txt	1.1088	27	0.5793	33
VISTA@UEvora	SVR-Run-01-sk-LDA.txt	1.1770	28	0.5918	27
HHU-DBS_2	SVR_RanFRST_depth_2_new.txt	1.2040	29	0.6484	17
SD VA HCS/UCSD	SVR9.csv	1.2153	30	0.6658	11
SD VA HCS/UCSD	SVRSubmission.txt	1.2153	31	0.6984	7
HHU-DBS_2	SVR_DTree_Features_Best_Bin.txt	1.3203	32	0.5402	36
HHU-DBS_2	SVR_DTree_Features_Best.txt	1.3203	33	0.5848	28
HHU-DBS_2	SVR_DTree_Features_Best_All.txt	1.3714	34	0.6750	9
MostaganemFSEI	SVR_mostaganemFSEI.txt	1.4207	35	0.5836	30
Middlesex University	SVR-Gao-April27.txt	1.5145	36	0.5412	35

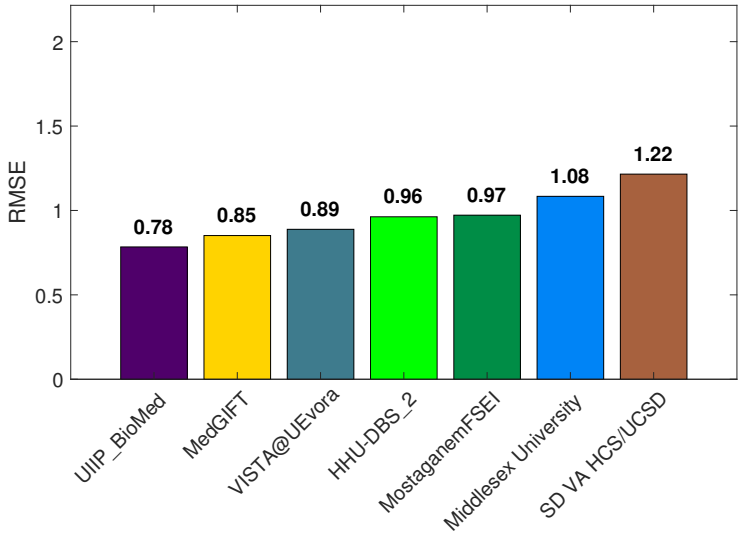

Fig. 6. Root Mean Square Error (RMSE) obtained by the best run of each group.

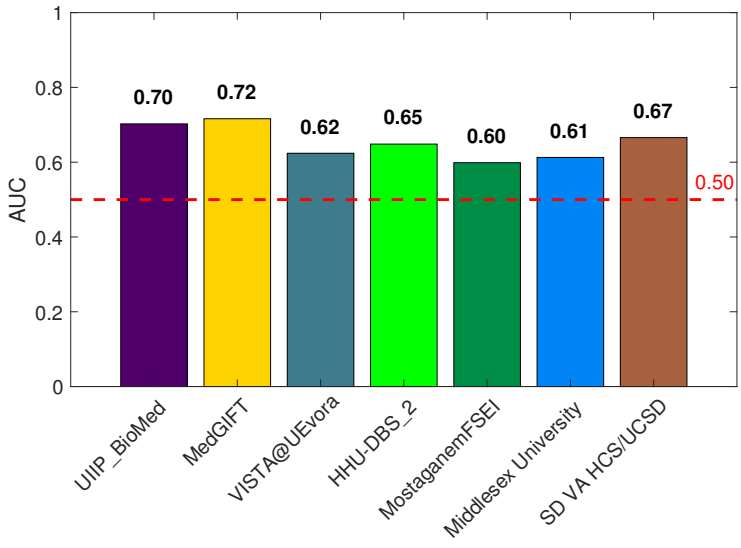

Fig. 7. Area Under the Curve (AUC) obtained by the best run (with respect to RMSE) of each group. The red line shows the AUC of a random classifier (0.50).

4 Discussion and Conclusions

Similar to 2017, the results obtained by the participants in the MDR detection subtask demonstrate that the task of a fully automatic image-based detection of

drug resistance is extremely difficult. Despite the addition of XDR TB cases into the dataset and the inclusion of information about patient age and gender, the MDR detection performance still remains at a level relatively close to a random classification with the highest reached AUC of 0.6178 and a 61.4% prediction accuracy. The overall increase of prediction performance with respect to the 2017 edition might be caused by the addition of more severe cases with XDR TB into the dataset. Using information about patient age and gender could also improve the MDR detection results as suggested by the baseline submitted by UniversityAlicante group [20].

In the second subtask, the overall results of TB type classification are slightly worse than in 2017. This might be caused by the decreased balance of TB classes in the dataset. Using more than one CT scan per patient could also confuse prediction methods and worsen the final results. However there is a certain improvement in prediction of class T2 (Focal TB) demonstrated by most of the participants' results.

The results of SVR subtask are encouraging, since the actual assessment of the TB severity score is done using various clinical information sources, not only CT image data. Most of the results achieved by the participants obtained a RMSE of the severity score below 1 in a 5-grade scoring system. The best results obtained using only CT volumes are close to the results reported in [28], where the authors used clinical and laboratory data including drug resistance, presence of TB symptoms, etc. in addition to the images. Extension of the dataset and usage of clinical and laboratory data is expected to improve the severity scoring results.

Overall, the 2018 edition of the ImageCLEF TB task showed an improvement with respect to the 2017 edition in terms of number of participants, data provided, results obtained and the variety of methods proposed. This shows a high interest in this topic and also the importance of the data that were generated.

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 DAA3-17-63599-1 "Year 6: 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., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. World Health Organization, et al.: Global tuberculosis report 2016. (2016)
11. 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)
12. 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
13. Liauchuk, V., Kovalev, V.: ImageCLEF 2017: Supervoxels and co-occurrence for tuberculosis CT image classification. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org <http://ceur-ws.org> (September 11-14 2017)

14. Wang, Y.X.J., Chung, M.J., Skrahin, A., Rosenthal, A., Gabrielian, A., Tartakovsky, M.: Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences. Quantitative Imaging in Medicine and Surgery **8**(2) (2018) 161–173
15. Kovalev, V., Liauchuk, V., Safonau, I., Astrauko, A., Skrahina, A., Tarasau, A.: Is there any correlation between the drug resistance and structural features of radiological images of lung tuberculosis patients? In: Computer Assisted Radiology - 27th International Congress and Exhibition (CARS-2013). Volume 8., Springer, Heidelberg (2013) 18–20
16. Kovalev, V., Liauchuk, V., Kalinouski, A., Rosenthal, A., Gabrielian, A., Skrahina, A., Astrauko, A., Tarasau: Utilizing radiological images for predicting drug resistance of lung tuberculosis. In: Computer Assisted Radiology - 27th International Congress and Exhibition (CARS-2015). Volume 10., Springer, Barcelona (2015) 129–130
17. Ahmed, M.S., Obaidullah, S.M., Jayatilake, M., Gonçalves, T., Rato, L.: Texture analysis from 3D model and individual slice extraction for tuberculosis MDR detection, type classification and severity scoring. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
18. Gentili, A.: ImageCLEF2018: Transfer learning for deep learning with CNN for tuberculosis classification. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
19. Tatusch, M., Conrad, S.: Detection of multidrug-resistant tuberculosis using convolutional neural networks and decision trees. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
20. Llopis, F., Fuster-Guilló, A., Rico-Juan, J.R., Azorín-López, J., Llopis, I.: Tuberculosis detection using optical flow and the activity description vector. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
21. Liauchuk, V., Tarasau, A., Snezhko, E., Kovalev, V., Gabrielian, A., Rosenthal, A.: ImageCLEF 2018: Lesion-based TB-descriptor for CT image analysis. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
22. Dicente Cid, Y., Müller, H.: Texture-based graph model of the lungs for drug resistance detection, tuberculosis type classification, and severity scoring: Participation in ImageCLEF 2018 tuberculosis task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
23. Allaouzi, I., Benamrou, B., Ben Ahmed, M.: 3D-CNN in drug resistance detection and tuberculosis classification. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
24. Ishay, A., Marques, O.: Ensemble of 3D CNNs with multiple inputs for tuberculosis type classification. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
25. Hamadi, A., Yagoub, D.E.: ImageCLEF 2018: Semantic descriptors for tuberculosis CT image classification. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)

26. Bogomasov, K., Himmelspach, L., Klassen, G., Tatusch, M., Conrad, S.: Feature-based approach for severity scoring of lung tuberculosis from CT images. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
27. Gao, X., James-Reynolds, C., Currie, E.: Scoring TB severity with an enhanced deep residual learning depth-resnet. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
28. Kovalev, V., Liauchuk, V., Skrahina, A., Astrauko, A., Rosenthal, A., Gabrielian, A.: Examining the utility of clinical, laboratory and radiological data for scoring severity of pulmonary tuberculosis. In: Computer Assisted Radiology and Surgury - 32nd International Congress and Exhibition (CARS-2018). Volume 13., Springer, Heidelberg (2018) 143–144