

ImageCLEF 2018: Lesion-based TB-descriptor for CT Image Analysis

Vitali Liauchuk¹, Aleh Tarasau², Eduard Snezhko¹, and Vassili Kovalev¹

¹ United Institute of Informatics Problems, Minsk, Belarus
vitali.liauchuk@gmail.com

² Scientific and Practical Center for Pulmonology and Tuberculosis, Minsk, Belarus

Abstract. The paper presents image description and classification method which was used by United Institute of Informatics Problems (UIIP_BioMed) group for accomplishing the three subtasks of ImageCLEFtuberculosis task. The image description method employed is based on automated detection of tuberculosis (TB) lesions of different types in 3D lung Computed Tomography (CT) scans. The lesion detection method is based on Coder-Decoder Convolutional Neural Network trained on a third-party dataset of 149 CT scans with lesions labeled by a qualified radiologist. It was shown that combination of lesion-based TB-descriptor and Random Forests classifier allows achieving the best performance in TB type classification and TB severity scoring subtasks.

Keywords: tuberculosis, TB-descriptor, lesions, CT, image analysis

1 Introduction

The tuberculosis task [3] of ImageCLEF 2018 Challenge [5] considers three subtasks all dealing with 3D CT images. The subtask #1 is dedicated to the problem of single image-based distinguishing between multi-drug resistant tuberculosis (MDR TB) cases and drug sensitive (DS) ones. The task remains very challenging and so far has no solution with sufficient prediction accuracy. Recent analysis of published evidences reports presence of statistically significant links between drug resistance and multiple thick-walled caverns [12]. So far computerized methods demonstrate performance of image-based detection of MDR TB barely beyond the level of statistical significance [4, 8, 9]. Compared to 2017 data [2], datasets for MDR detection subtask were extended by means of adding several cases with extensively drug-resistant tuberculosis (XDR TB), which is a rare and more severe subtype of MDR TB. Thus, training data for the MDR detection subtask included 259 CT images: 134 drug sensitive and 125 drug resistant cases. Test set consisted of 236 CT images: 101 drug sensitive and 135 drug resistant cases.

The subtask #2 of ImageCLEFtuberculosis task is aimed at automatic categorization of CT images into one of five types of tuberculosis: Infiltrative, Focal, Tuberculoma, Miliary and Fibro-cavernous. Compared to 2017, the datasets were extended by adding new CT scans of the patients involved earlier, and also by

introducing CT images of some new patients. However, in this study only the first CT scan of each patient was used.

The newly represented subtask #3 was dedicated to assessment of severity of TB based on a single CT image of a patient. The severity score has meaning of a cumulative score of severity of TB case assigned by a medical doctor. Originally, the severity scores were assigned using natural numbers between 1 ("critical/very bad") and 5 ("very good"). Additionally, for the case of binary classification the scores were converted to binary values where scores from 1 to 3 corresponded to "high severity" and the remaining 4 and 5 corresponded to "low severity". In the process of scoring, the medical doctors considered many factors like patterns of lung lesions, results of microbiological tests, duration of treatment, patient's age and some other. One of the goals of this subtask is to distinguish "low severity" from "high severity" based solely on the CT scan.

2 Detection of lung lesions in CT, TB-descriptor

In this section, a method for automated detection of lung lesions in 3D CT images is described. The method is based on training the Deep Convolutional Neural Network (CNN) on a set of data derived from 3D CT images with manually labeled lesions of different types. The method utilizes slice-wise image segmentation technique previously described in [6]. This technique considers splitting the original 3D image into a number of smaller 2D regions, processing the regions one-by-one and collecting the CNN output into a 3D probability map (see Fig. 1). Finally, a quantitative TB-descriptor is built based on the lesion probability maps.

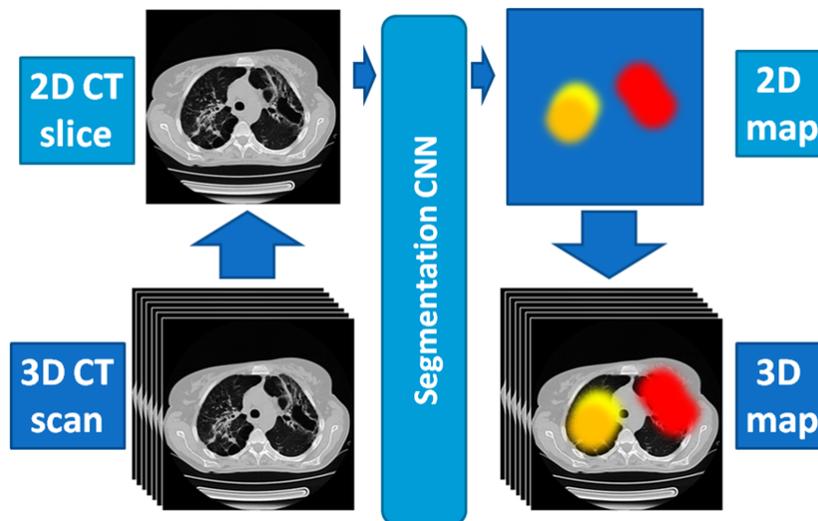


Fig. 1. General scheme of the slice-wise lesion segmentation method

2.1 Data preparation

TB lesions were labeled manually on a total number of 198 3D CT scans. The labeling was performed in two stages. The first stage was performed by a qualified radiologist and was aimed at coarse localization of TB lesions of different type in lungs without the exact delineation. The second stage was aimed at correction of initial lesion labeling and making more precise segmentation of lesions (see Fig. 2). Both stages of labeling were performed using an auxiliary software tool designed by the authors (see Fig. 3).

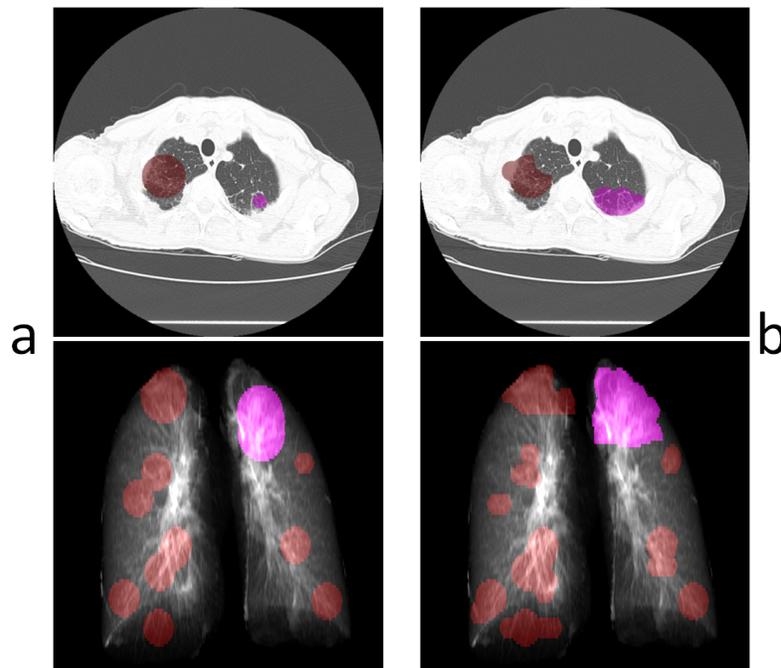


Fig. 2. Labeling stages, axial slices (top) and frontal projections (bottom): a) initial stage, rough labeling; b) second stage, more precise segmentation of lesions

The developed software tool allows labeling of 10 different types of TB lesions. Some types of lesions were well represented in the dataset whilst lesions of some other types (Plevritis, Atelectasis, Pneumathorax) were present only in few images in the dataset. List of lesion types and the corresponding frequencies of occurrence in dataset images are shown in Table 1. In the result of labeling process, 3D masks with the corresponding lesion indexes were obtained.

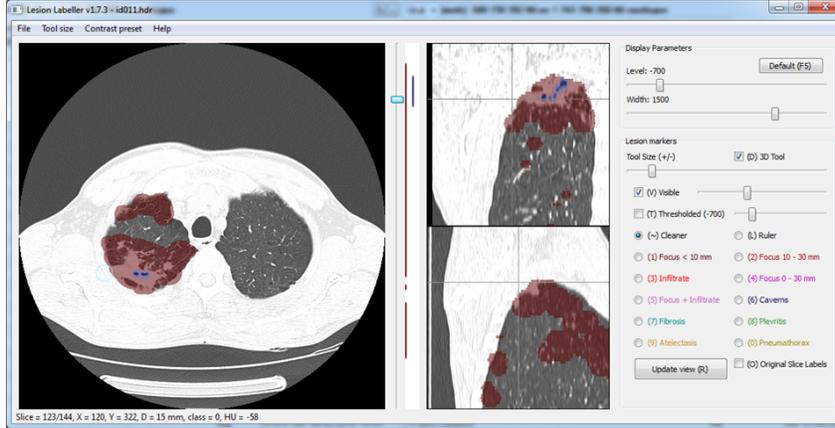


Fig. 3. Screenshot of the developed software tool for lesion segmentation

Table 1. Presence of lesions of different types in the dataset

Index	Type of lesion	Number of images
1	Focus < 10 mm	140
2	Focus 10–30 mm	38
3	Infiltrate	26
4	Focus 0–30 mm (mix)	85
5	Focus + Infiltrate (mix)	30
6	Caverns	81
7	Fibrosis	56
8	Plevritis	13
9	Atelectasis	7
10	Pneumathorax	4

2.2 Segmentation of lung regions

For extraction of lung regions for both lesion detection and ImageCLEFtuberculosis subtasks, a domestic implementation of a conventional segmentation-by-registration approach [11] was employed instead of the one proposed by the organizers. In our case the method utilized 130 reference CT scans with manually segmented lungs. Projections along X , Y and Z axes are calculated for each reference CT scan. The three normalized projections are concatenated into a quantitative descriptor of a reference image. For a target CT scan, a similarity measure is calculated between the target image and the reference images based on the quantitative descriptors of all images. Top-5 most similar reference images are selected. The selected images along with the corresponding lung masks are non-rigidly registered to the target image using 'elastix' software tool [7], final segmentation mask is obtained by means of averaging. The implemented method demonstrates high robustness to the presence of large lesion in lungs (see Fig. 4).

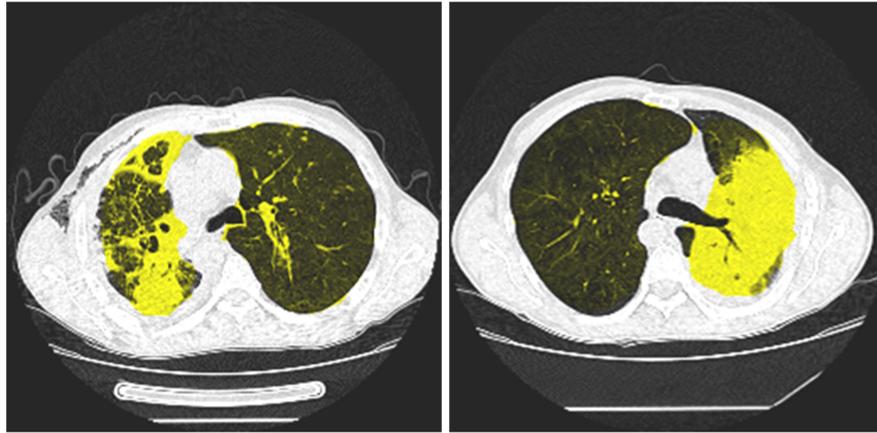


Fig. 4. Example slices of CT images with segmented lungs

2.3 Training the Convolutional Neural Network

One of the possible ways to employ Deep Learning algorithms for 3D image is to operate at slice level by representing each 3D CT image as a set of 2D slices. One of the advantages of such approach is relatively low usage of computer memory since the large 3D is processed slice-by-slice. In the current study, 2D image regions of size 128×128 pixels were extracted from slices of original CT images with 64-pixels stride. Three neighboring slices were used to compose a single RGB image in order to use spatial information along Z-axis of original CT images. Finally, the image regions were up-sized using bicubic interpolation to 256×256 pixels. The up-sizing was performed to improve the detection of small lesions since the first convolutional layer of the network used which is AlexNet has 4-pixel stride, and some lesions present on the images have size of 2–3 pixels.

From the total amount of 198 labeled 3D scans, 149 were used for training the algorithms and the rest 49 were used for validation. Lesion types with indexes 1–5 were merged together into one class "Foci" as having similar nature and/or being mixture of classes. From the 149 training CT images, 268,278 2D image tiles were extracted. For each tile a corresponding label image was composed using manually labeled lesion data (see Fig. 5). Image regions which lay beyond the lung segmentation masks are marked with a special "don't care" label. Neural network omits these regions at both training and validation stages which allows to better focus the available computational facilities on the actual regions of interest. On the label images such regions are marked with gray color.

For segmentation of lesions in 2D slice regions a Fully Convolutional Network Alexnet [10] was used. In order to increase convergence rate and overall accuracy, a publicly available ILSVRC2012-trained model was used to initialize the networks weights. The net was set to recognize multiple lesion types at the same time.

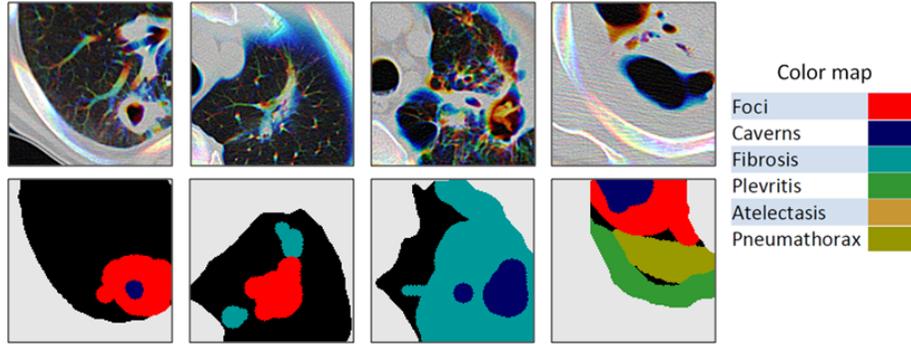


Fig. 5. Examples of 2D slice regions (top row) and the corresponding label images (bottom row)

Training was performed on a personal computer equipped with Intel i7-6700K CPU and dedicated GPU of Nvidia TITAN X type with 3072 CUDA Cores and 12 GB of GDDR5 onboard memory. NVIDIA DIGITS interface and Caffe framework were used. The network training parameters were set to the following values: Number of epochs=60, Activation function=ReLU, Batch size=64, Solver type=SGD Caffe solver. Learning Rate was set to 0.001 for the first 20 epochs, 0.0001 for the next 20 and 0.00001 for the last 20 ones.

2.4 Obtaining probability maps

Once the training process is finished, the trained network model can be used for detection of lesions in an arbitrary 3D CT scan. In this case the CT image undergoes the same procedures as for the training images:

- segmentation of lung regions;
- extraction of 2D tiles;
- processing the tiles with the trained CNN and obtaining probability maps for each lesion type considered;
- collecting the obtained 2D probability maps into 3D probability maps for each lesion type separately;

Additionally, probability maps can be smoothed to reduce the number of falsely detected lesions in images, or thresholded so that all probability values below minimum allowed value are zeroed. Fig. 6 demonstrates the detected lesions on test CT scans. Lesion regions were obtained from the corresponding probability maps by means of thresholding with $P_{thres} = 0.5$. The resultant lesion regions are marked with colors with correspondence to the colormap from Fig. 5.

2.5 Building TB-descriptor

Once the probability maps are built, the TB-descriptor proposed with this study is built as follows. The lungs region on CT image is divided into 6 parts as it is

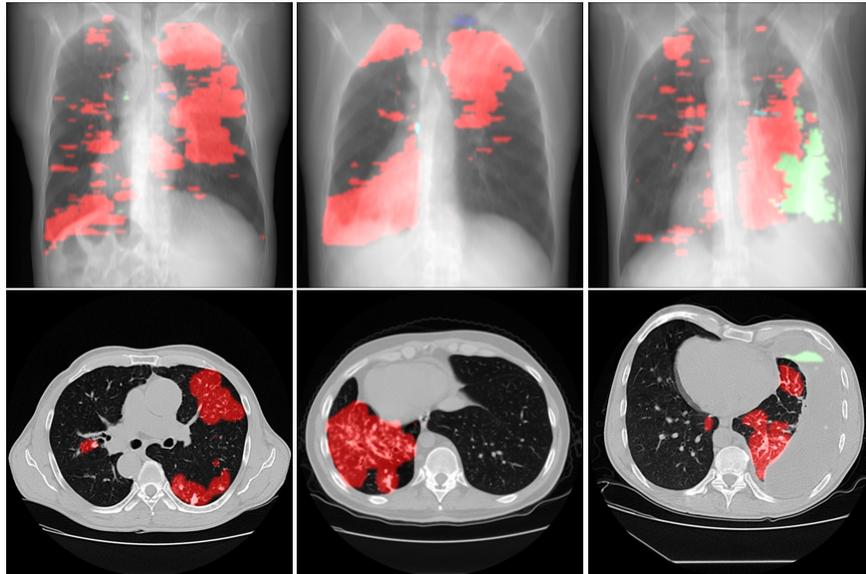


Fig. 6. Detected lesions on test CT images: frontal projections (top row) and axial slices (bottom row)

shown on Fig. 7. Height of the parts along Z axis is taken equal. For every type of lesion its presence in each of six parts is calculated as the sum of probabilities in the corresponding voxels divided by the number of lung voxels within the considered part. Since all the probabilities are ranged from 0 to 1, the lesion presence score for each part is also a number from 0 to 1. Finally, the presence scores obtained for each lesion type and each lung part are concatenated into a single TB-descriptor of size $N_{lesion_types} \times N_{parts}$.

Thus, the proposed TB-descriptor indicates presence of lesions of certain types in different parts of lungs: upper left, middle right, etc. Portion of the affected lung volume is considered as well. Such TB-descriptor was used for recognition of drug resistance status, type and severity of tuberculosis in the ImageCLEF challenge subtasks.

3 Submissions and results

For all the ImageCLEF tuberculosis subtasks the following prediction scheme was used:

- segmentation of lung regions for each CT image;
- detection of lesions;
- calculation of TB-descriptors for each image;
- prediction of the desired values using a valid classifier.

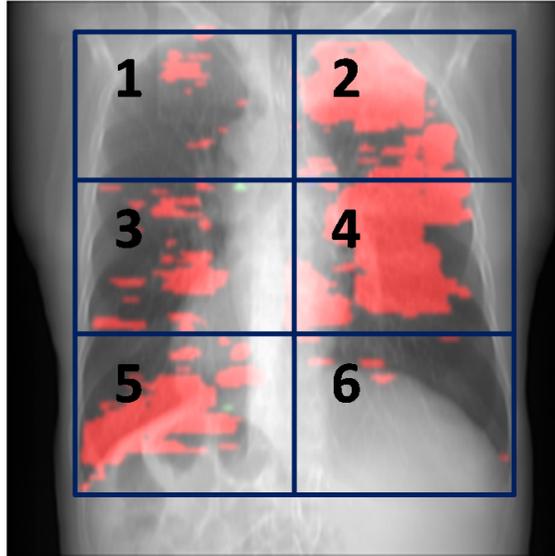


Fig. 7. Lungs region partitioning

Subtasks of ImageCLEFtuberculosis considered different types of predictions: multiple-class prediction where only the index of predicted class must be provided, two-class prediction where probability of belonging to positive class must be provided as well, and regression where the corresponding method needs to predict value of a continuous variable as precise as possible. For all three subtasks, Random Forests classifier was used which is capable of handling all the above-mentioned tasks. Assessment of the algorithms performance was carried out on the Training data using k -fold cross-validation procedure with $k = 5$.

3.1 Subtask #1: MDR detection

Following the above-mentioned prediction scheme, TB-descriptors were calculated for all the available CT images. Random Forests classifier was trained on the set of TB-descriptors with concatenated meta-data values: patients' age and gender. Based on a series of experiments, number of trees in the classifier was chosen to be 150 for this subtask. Accuracy assessment within 5-fold cross-validation demonstrated Area Under ROC-Curve (AUC) value of 0.6385. One run was submitted as the result of prediction of test data.

A total number of 39 runs were submitted by 7 different participating groups for MDR detection subtask. Table 2 shows top-15 best participants' results in terms of AUC value. Utilizing lesion-based TB-descriptor resulted in 0.5558 AUC and ranked 14-th place among the 39 runs. The best achieved result by VISTA@UEvora team with 0.6178 AUC value outperforms previous year's result with 0.5825 AUC. However, MDR detection performance still remains at a level close to random classification. Increase of prediction performance might be

caused by adding a number of more severe cases with XDR TB into the dataset and also by utilizing information about patients’ age and gender.

Table 2. Top-15 submitted runs with highest AUC values for MDR detection subtask.

Group Name	Run	AUC	Rank
VISTA@UEvora	06-Mohan-SL-F3-Personal	0.6178	1
San Diego VA HCS/UCSD	MDSTest1a	0.6114	2
VISTA@UEvora	08-Mohan-voteLdaSmoF7-Personal	0.6065	3
VISTA@UEvora	09-Sk-SL-F10-Personal	0.5921	4
VISTA@UEvora	10-Mix-voteLdaSI-F7-Personal	0.5824	5
HHU-DBS	FlattenCNN_DTree	0.5810	6
HHU-DBS	FlattenCNN2_DTree	0.5810	7
HHU-DBS	Conv68adam_fl	0.5768	8
VISTA@UEvora	07-Sk-LDA-F7-Personal	0.5730	9
UniversityAlicante	MDRBaseline0	0.5669	10
HHU-DBS	Conv48sgd	0.5640	11
HHU-DBS	Flatten	0.5637	12
HHU-DBS	Flatten3	0.5575	13
UIIP_BioMed	TBdescs2_zparts3_thrprob50_rf150	<u>0.5558</u>	<u>14</u>
UniversityAlicante	testSVM_SMOTE	0.5509	15

3.2 Subtask #2: TBT classification

For TB type classification subtask, a similar procedure was carried out with the difference that Random Forests classifier was trained for the case of multiple image classes. Number of trees for this subtask was chosen to be 150. Instead of using all the available data, only the first CT scan of each patient was used both for algorithms training and for final prediction of patient’s TB class.

In total, 39 runs were submitted by 8 participating groups for TB type classification subtask. The results were evaluated and ranked by accuracy and Cohen’s Kappa coefficient [1] which is preferable in the case of unbalanced dataset. Among the submitted runs our method based on lesion detection demonstrated the best TB type recognition performance in terms of both Kappa coefficient (0.2312) and accuracy (0.4227) (see Table 3). Compared to 2017, overall TB type classification results are less accurate. Probably this is caused by the increased disbalance between TB types. Using more than one CT scan per patient might also confuse prediction methods and worsen the final results.

3.3 Subtask #3: Severity scoring

In contrast to the two previous subtasks, the TB severity scoring subtask was evaluated in two principally different ways.

One way of evaluation used the original severity scores from 1 to 5 as provided by the doctors and the task for participants was to predict those numerical scores

Table 3. Top-15 submitted runs with highest Kappa values for TB type subtask.

Group Name	Run	Kappa	Rank
<u>UIIP_BioMed</u>	<u>TBdescs2_zparts3_thrprob50_rf150</u>	<u>0.2312</u>	<u>1</u>
fau_ml4cv	m4_weighted	0.1736	2
MedGIFT	AllFeats_std_euclidean_TST	0.1706	3
MedGIFT	Riesz_AllCols_euclidean_TST	0.1674	4
VISTA@UEvora	02-Mohan-RF-F20I1500S20-317	0.1664	5
fau_ml4cv	m3_weighted	0.1655	6
VISTA@UEvora	05-Mohan-RF-F20I2000S20	0.1621	7
MedGIFT	AllFeats_AllCols_correlation_TST	0.1531	8
MedGIFT	AllFeats_mean_euclidean_TST	0.1517	9
MedGIFT	Riesz_std_euclidean_TST	0.1494	10
San Diego VA HCS/UCSD	Submission64a	0.1474	11
San Diego VA HCS/UCSD	TBTTask_2_128	0.1454	12
MedGIFT	AllFeats_AllCols_correlation_TST	0.1356	13
VISTA@UEvora	03-Mohan-RF-7FF20I1500S20-Age	0.1335	14
San Diego VA HCS/UCSD	TBTLast	0.1251	15

as precise as possible. Here, Root Mean Square Error (RMSE) was computed between ground truth and predicted severity scores provided by participants. The goal was to achieve lowest possible RMSE value.

The other way of evaluation considered binary classification problem. The original severity index was transformed into two class values: cases with scores from 1 to 3 were labeled as "high severity" cases and the other cases with scores 4 and 5 corresponded to "low severity" class. With this way of evaluation the participants were to provide probabilities of TB cases belonging to "high severity" class. The results were ranked using AUC value. Top-10 runs for both evaluation methods are shown in Tables 4 and 5.

Table 4. Top-10 submitted runs with lowest RMSE values for Severity scoring.

Group Name	Run	RMSE	Rank
<u>UIIP_BioMed</u>	<u>TBdescs2_zparts3_thrprob50_rf100</u>	<u>0.7840</u>	<u>1</u>
MedGIFT	HOG_std_euclidean_TST	0.8513	2
VISTA@UEvora	07-Mohan-MLP-6FTT100	0.8883	3
MedGIFT	AllFeats_AllCols_euclidean_TST	0.8883	4
MedGIFT	AllFeats_AllCols_correlation_TST	0.8934	5
MedGIFT	HOG_mean_euclidean_TST	0.8985	6
MedGIFT	HOG_mean_correlation_TST	0.9237	7
MedGIFT	HOG_AllCols_euclidean_TST	0.9433	8
MedGIFT	HOG_AllCols_correlation_TST	0.9433	9
HHU-DBS	RanFrst	0.9626	10

In total, 36 runs were submitted by 7 participants for this subtask. As it can be seen from the tables, lesion-based TB-descriptor appeared to be extremely

Table 5. Top-10 submitted runs with highest "low severity"/"high severity" prediction performance.

Group Name	Run	AUC	Rank
MedGIFT	AllFeats_AllCols_correlation_TST	0.7708	1
MedGIFT	HOG_AllCols_correlation_TST	0.7608	2
MedGIFT	HOG_mean_euclidean_TST	0.7443	3
MedGIFT	HOG_AllCols_euclidean_TST	0.7268	4
MedGIFT	HOG_std_euclidean_TST	0.7162	5
<u>UIIP_BioMed</u>	<u>TBdescs2_zparts3_thrprob50_rf100</u>	<u>0.7025</u>	<u>6</u>
San Diego VA HCS/UCSD	SVRSubmission	0.6984	7
HHU-DBS	RanFRST_depth_2_Ludmila_new_new	0.6862	8
HHU-DBS	DTree.Features_Best_All	0.6750	9
MedGIFT	AllFeats_AllCols_euclidean_TST	0.6733	10

useful for assessing TB severity with the best result in terms of regression (minimum RMSE among all runs) and 6-th best result in terms of "low severity"/"high severity" classification. Number of trees for this experiments was set to 100. The highest binary classification performance with AUC value of 0.7708 was achieved by MedGIFT group.

4 Conclusions

The results of this study allows to draw the following conclusions:

- Combination of lesion-based TB-descriptor and Random Forests classifier allowed achieving the best performance in TB type classification and TB severity scoring subtasks.
- Similar to 2017 results, image-based MDR TB detection performance remains low (AUC 0.6178, accuracy 55.93%) despite the addition of XDR TB cases into the dataset and utilizing information about patients' age and gender.
- Lesion-based TB-descriptor derived from lung CT scans conveys valuable information on patient's state and is worth to consider in CT image analysis of TB patients.
- Extending the training data for lesion detection is desirable for further improvements of computerized TB diagnosis.

In this paper, image description and analysis method based on automatic detection of TB lesions in lungs and composing TB-descriptor is presented. The method was employed by UIIP_BioMed group in all three subtasks of Image-CLEFtuberculosis 2018 challenge.

Acknowledgements

This study was supported 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. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46 (1960)
2. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Dublin, Ireland (September 11-14 2017)
3. Dicente Cid, Y., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Avignon, France (September 10-14 2017)
4. 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. Lecture Notes in Computer Science, vol. 10456. Springer, Dublin, Ireland (September 11-14 2017)
5. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andreczyk, 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), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
6. Kalinovsky, A., Liauchuk, V., Tarasau, A.: Lesion detection in CT images using Deep Learning semantic segmentation technique. In: International Workshop "Photogrammetric and computer vision techniques for video surveillance, biometrics and biomedicine". The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII, pp. 13–17. Moscow, Russia (May 2017). <https://doi.org/10.5194/isprs-archives-XLII-2-W4-13-2017>, <http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W4/13/2017/>
7. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on medical imaging* **29**(1), 196–205 (2010)
8. 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). vol. 10, pp. 129–130. Springer, Barcelona (2015)
9. 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). vol. 8, pp. 18–20. Springer, Heidelberg (2013)

10. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 640–651 (April 2017). <https://doi.org/10.1109/TPAMI.2016.2572683>
11. Sluimer, I., Prokop, M., van Ginneken, B.: Toward automated segmentation of the pathological lung in ct. *IEEE Transactions on Medical Imaging* **24**(8), 1025–1038 (Aug 2005). <https://doi.org/10.1109/TMI.2005.851757>
12. Wang, Y.X.J., Chung, M.J., Skrahin, A., Rosenthal, A., Gabrielian, A., Tarkovsky, M.: Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences. *Quantitative Imaging in Medicine and Surgery* **8**(2), 161–173 (2018)