ImageCLEF2019: Tuberculosis - Severity Scoring and CT Report with Neural Networks, Transfer Learning and Ensembling

Amilcare Gentili[1-2][0000-0002-5623-7512]

[1] San Diego VA Health Care System, San Diego, CA USA
[2] University of California, San Diego, CA, USA
agentili@ucsd.edu

Abstract. The diagnosis of tuberculosis is challenging. We present our approach for classifying whether a patient has high or low severity tuberculosis and for detecting which lung is involved, if there is decreased capacity, and if there are pleurisies, calcifications or cavities present. Our best results for the CT report task were obtained by converting volume images into an 8x4 montage of sagittal or coronal images and ensembling the results of separate networks trained separately on sagittal and coronal montage images. The best results for the severity scoring were obtained by ensembling the results from the CT report with the provided metadata.

Keywords: Deep Learning, Convolutional Neural Network, Tuberculosis, CT Scans.

1 Introduction

Tuberculosis is a common disease where fast diagnosis using CT images can often improve treatment results. An accurate and automatic method for classifying tuberculosis from CT images may be especially useful in regions of the world with few radiologists. The ImageCLEF 2019[1] has 2 challenges [2]: 1) scoring severity of tuberculosis from CT images and 2) creating a report that identifies if the left lung is affected, if the right lung is affected, if calcifications, caverns, and/or pleurisy are present, and if lung capacity is decreased.

2 Methods

2.1 Data.

The data set provided for both the CT report subtask and severity scoring subtask of the ImageCLEF 2019 Tuberculosis task [2] use the same dataset containing 335 chest CT scans of TB patients along with a set of clinically relevant metadata. 218 patients are used for training and 117 for test. The provided metadata includes information about

disability, relapse, symptoms of TB, comorbidity, bacillary, drug resistance, education level, incarceration history, alcohol consumption, and smoking history. A set of lung masks was also provided for all patients[3] .

For the CT report task, the training set distribution of pathology was somewhat unbalanced with lung involvement being very common, and calcifications and pleurisy rare.

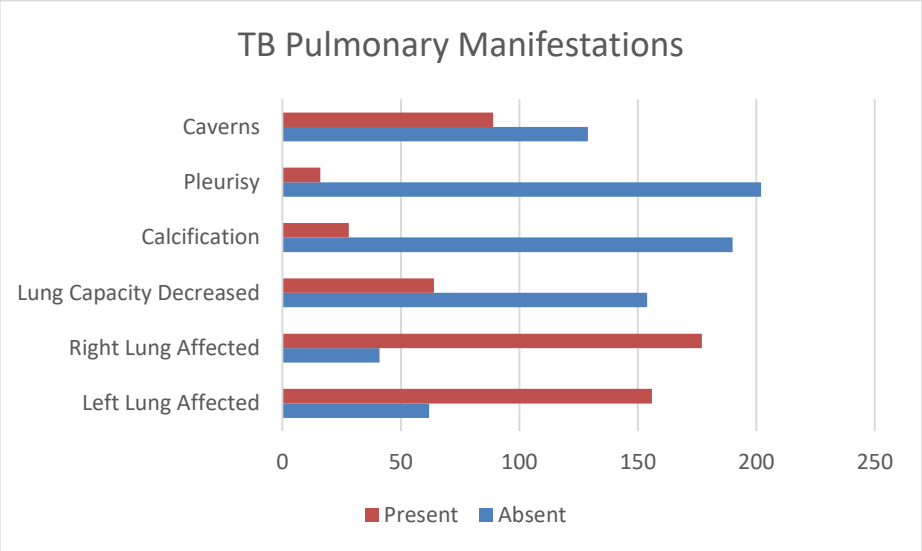

Fig. 1. Distribution of manifestations of tuberculosis in the training dataset

For the severity scoring task, the training set distribution of high and low severity was balanced. See Figure 2

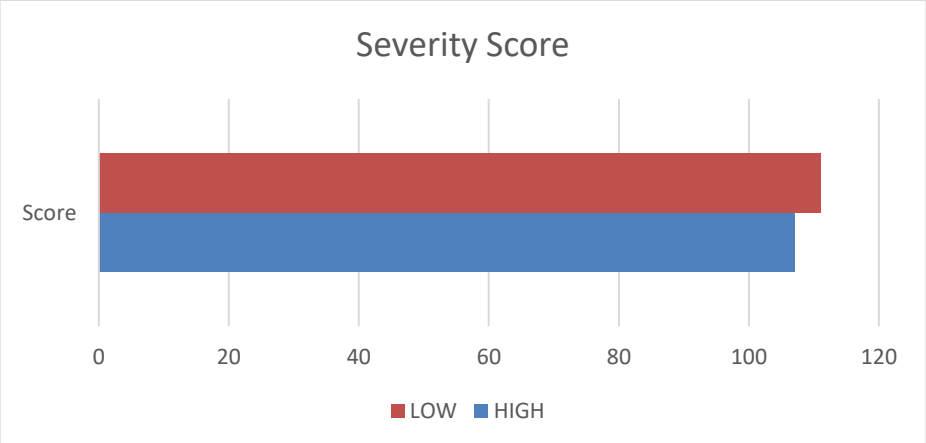

Fig. 2. Distribution of high and low severity score in the training dataset

2.2 Metadata Analysis

Reviewing the metadata shows that some factors are a strong predictor of high severity score. See Table 1

Table 1. Odd ratio of high severity for different factors

Factor	High	Low	Total	OR
Comorbidity	71	51	122	2.32
Disability	25	9	34	3.46
Symptoms Of TB	69	48	117	2.38
Relapse	50	26	76	2.87
Drug Resistance	88	51	139	5.45
Bacillary	99	86	185	3.60
Higher Education	7	21	28	0.30
Alcoholic	30	19	49	1.89
Ex-Prisoner	19	8	27	2.78
Smoking	63	51	114	1.68
Left Lung Affected	89	67	156	3.25
Right Lung Affected	94	83	177	2.44
Lung Capacity De-crease	43	21	64	2.88
Calcification	14	14	28	1.04
Pleurisy	14	2	16	8.20
Caverns	58	31	89	3.05

Drug resistance, disability, and bacillary had the strongest influence on increasing probability of high severity, and higher education had the strongest influence on increasing probability of low severity.

2.3 Preprocessing.

The images for the ImageCLEF tuberculosis task were provided as NIfTI 3D datasets. We used two different approaches for preprocessing images. For the first run (SVT_5, CTR_3) we used a method similar to what we employed for the ImageCLEF 2018 challenge [4]. We converted the images using med2image, a Python3 utility that converts medical image formatted files to more visual friendly ones, such as png and jpg. After reconstructing them in all three planes, we decided to use the coronal plane images, since they had the most images containing areas of abnormal lung. Although we did not visually verify the images of this data set, tuberculosis usually involves the upper lobes with relatively unaffected lung bases. As a result, axial images through the lung bases could possibly be normal even in patients with severe disease in the upper lobes. As med2image did not take in consideration slice thickness, the reconstructed coronal images were deformed and of different height. To correct this problem, all images were resized to a 512 x 512 matrix. Image masks for the lungs were available[3], and were used to select the 200 images with the largest area of lung in the image. For the first

run, all image equalization and data augmentation was done at the time of training using the fastai library [5].

For further runs (SVR_1, SVR_2, SVR_3, CTR_1, CTR_2 we used a different approach. We use nibabel library [6] to convert the NIfTI 3D datasets into numpy 3D arrays, using the provided lung masks [3], we cropped the 3D arrays to the smallest parallelogram that includes mostly the lungs. We equalized the array. We reshaped the array to have 31-32 slices in either the sagittal or coronal plane with a 256x256 matrix. Using montage, we combined the images into a single image. We did not correct for difference in slice thickness. See Figure 1 and 2. Data augmentation was done at the time of the training using the fastai library.

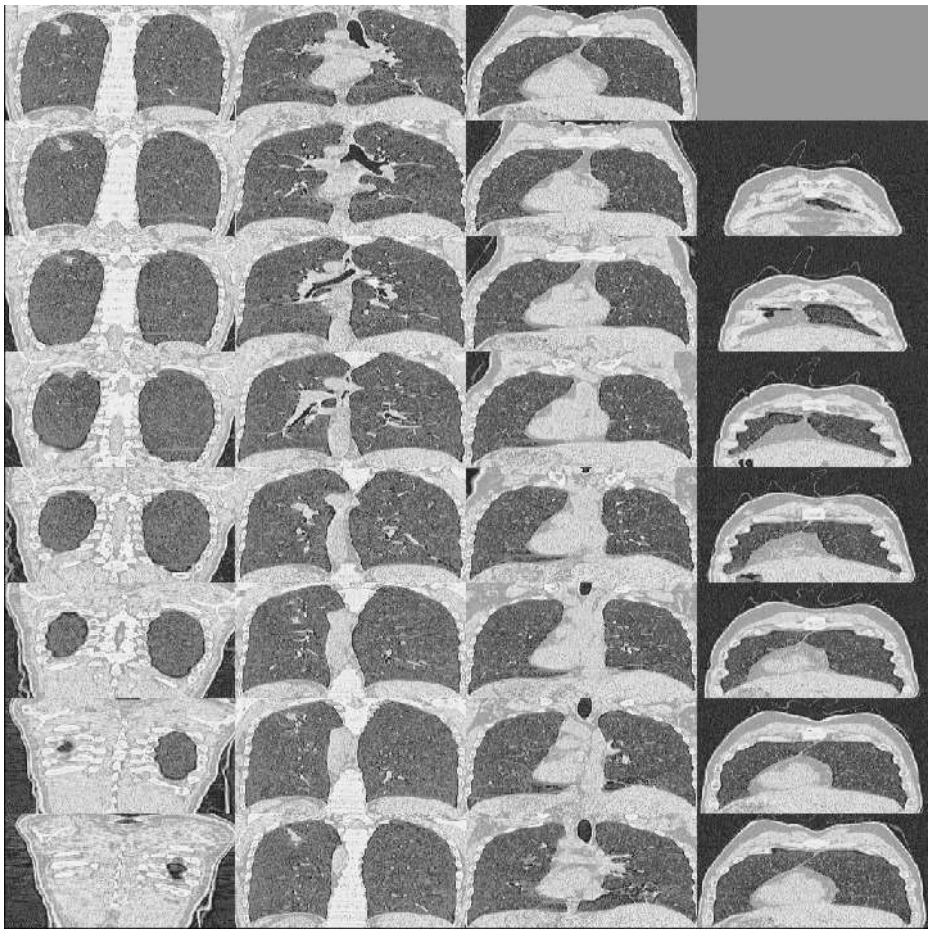

Fig. 3. Montage of equalized images in the coronal plane.

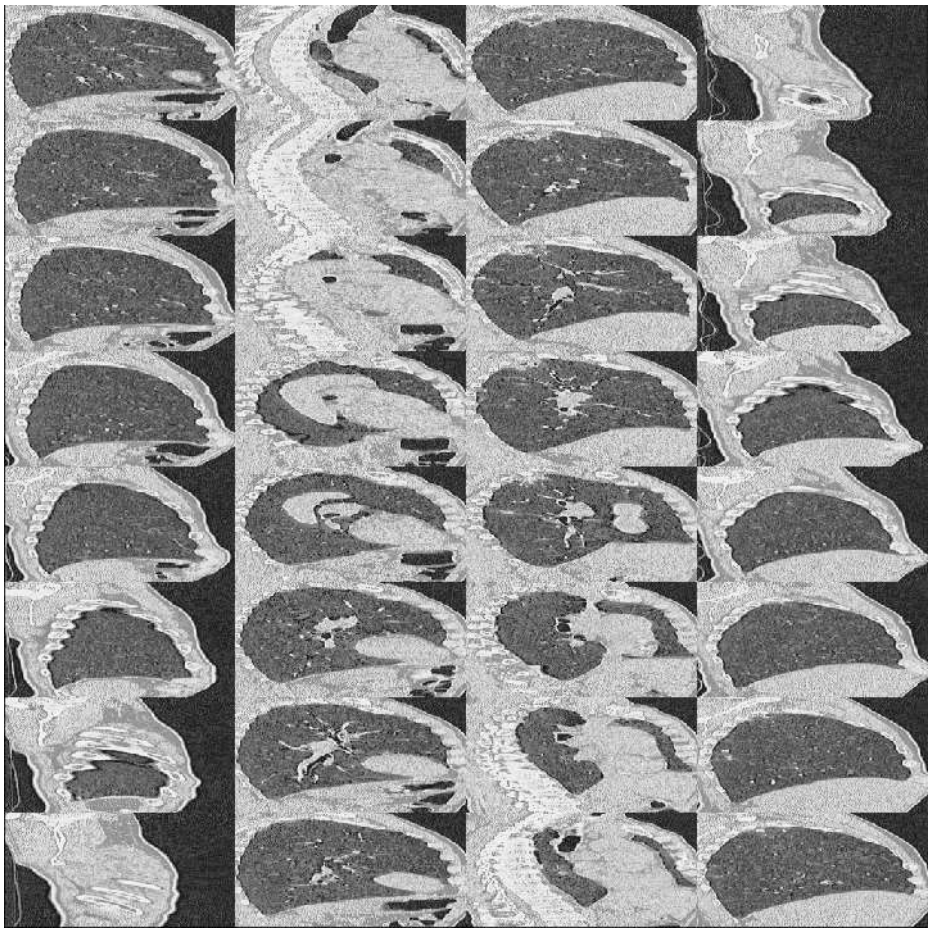

Fig. 4. Montage of equalized images in the sagittal plane.

2.4 Neural Network Training

For training the neural network, we used a workstation with an AMD Ryzen Threadripper 1950X CPU with 16 CPU cores and 32 threads, a Nvidia Quadro P6000 GPU, 64 GB RAM, and a 1 TB solid state drive. We took advantage of the fastai library to perform transfer learning of convolutional neural networks. We tried the following architectures that were available in the fastai library: resnet18, resnet34, resnet50, resnet101, resnet152, squeezenet1_0, squeezenet1_1, densenet121, densenet161, densenet169, densenet201, vgg16_bn, vgg19_bn, and alexnet. Resnet50, resnet101, densenet121, densenet161, and densenet169 gave the best results, so we decided to ensemble them.

For training the CNN, image sizes of 224x224, 299x299, and 384x384 were utilized. The learning rate was determined after running the learning rate finder function and plotting the learning rate vs. loss.

2.5 Ensembling results and metadata analysis

Orange [7] was used to create a prediction based on metadata only (SVR_4), and to combine metadata results with neural network results (SVR_1, SVR_2). See Figure 5.

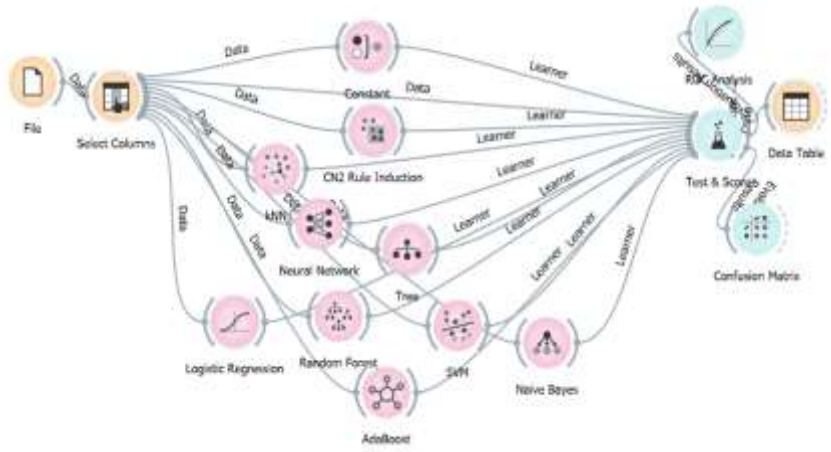

Fig. 5. Example of Orange 3 workflow to compare different machine learning approaches.

3 Results

3.1 CT Report Task

For the CTR_3 submission, for each patient we took the 200 images with the largest lung surface, scored each of those images separately using all pre-trained CNNs available in the fastai library, and averaged those results. Both mean AUC and minimum AUC were low, probably because only a few images of each patient have pathology, and averaging results decreased the probability of positive results.

For the CTR_1 and CTR_2 submissions we created a 4x8 montage of sagittal or coronal images for each patient. W separately scored sagittal and coronal images with 6 neural networks. For the CTR_2 submission, we ensembled all results, and for the CTR_1 submission, we ensembled the 3 best results.

Table 2. CT Report Task

Run Id	Run	Mean AUC	Min AUC
CTR_1	CTR_Cor_32_montage.txt	0.6631	0.5541
CTR_2	CTR_ReportsubmissionEnsemble2.csv	0.6532	0.5904
CTR_3	TB_ReportsubmissionLimited1.csv	0.5811	0.4111

3.2 Severity Scoring Task

Table 3. Severity Scoring Task

Run Id	Run	AUC	Accuracy
SVR_1	SVR_From_Meta_Report1c.csv	0.7214	0.6838
SVR_2	SVR_Meta_Ensemble.txt	0.7123	0.6667
SVR_3	SVR_LAstEnsembleOfEnsemblesReportCl.csv	0.7038	0.6581
SVR_4	SVRMetadataNN1_UTF8.txt	0.6956	0.6325
SVR_5	SVT_Wisdom.txt	0.627	0.6581

For the SVR_5 submission, we once again took the 200 images with the largest lung surface of each patient. For each patient, we scored each of those 200 images separately using all pretrained neural networks available in the fastai library and averaged those results. Both AUC metrics were low, for similar reasons to the CT Report Task.

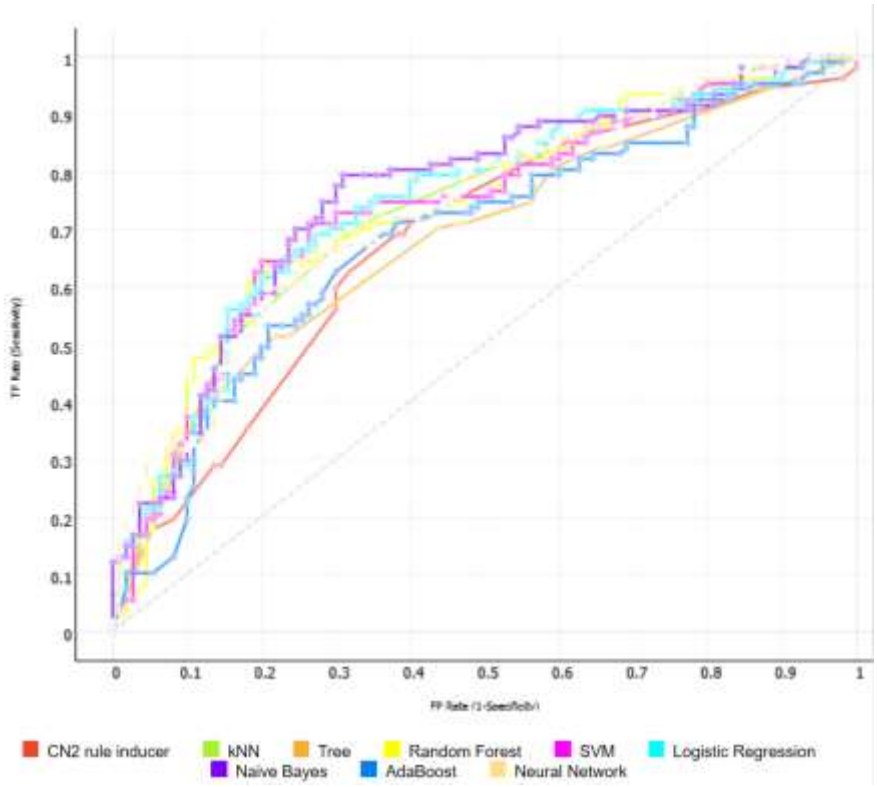

Fig. 6. ROC curves of different models trained using only the metadata of the training set, based on 10-fold cross validation, calculated with Orange3 workflow from Figure 3

For the SVR_4 submission, we trained different machine learning models available in Orange3 (Constant, AdaBoost, Tree, CN2 rule inducer, Random Forest, SVM, kNN,

Logistic Regression, Neural Network, Naive Bayes) and based on validation results we selected the top 4 to ensemble for the submission. See Figure 6.

For SVR_3 we took the results of classifying 4x8 montages of sagittal or coronal images as high or low severity, and ensembled them. For each 4x8 montage, we scored each sagittal and coronal image separately by ensembling the results of 6 neural networks.

For SVR_2 we ensembled SVR_3 with the metadata.

For SVR_1 we used Orange3 to create a model from the metadata using (Comorbidity, Disability, Symptoms of TB, Relapse, Drug Resistance, Bacillary, Higher Education, Alcoholic, Ex-Prisoner, Smoking) and training data (Left Lung Affected, Right Lung Affected, Lung Capacity Decrease, Calcification, Cavity, Pleurisy) and for the prediction we used the test metadata and the results from CTR_1 (Left Lung Affected, Right Lung Affected, Lung Capacity Decrease, Calcification, Cavity, Pleurisy). Although we tried Constant, AdaBoost, Tree, CN2 rule inducer, Random Forest, SVM, kNN, Logistic Regression, Neural Network, and Naive Bayes models, after evaluating the validation results, we used only SVM, Logistic Regression, Neural Network and Naive Bayes models to ensemble for the final submission.

4 Conclusion

In this paper, we presented the use of transfer learning to quickly train a CNN to classify the severity of tuberculosis and different pathological manifestations of tuberculosis.

5 Perspectives for Future Work

The training data set for the CT Report was imbalanced with only a few cases of calcification or pleurisy, but we did not try to compensate for this imbalance. Trying to compensate for this imbalance may improve results. We trained the neural network as a multilabel task on the same set of equalized images. Using images with different windows to enhance calcifications, training neural networks to detect just calcifications or just cavities, and using windows set to visually enhance air within the lungs, may improve results. Using Hounsfield units from the original images instead of values in the png files may also be more accurate. As our best results for the Severity Task came from combining the results of the CT Report Task with the metadata, improving results of the CT Report should improve results for the Severity Task too.

References

1. Ionescu, B., H. Müller, R. Péteri, Y.D. Cid, V. Liauchuk, V. Kovalev, D. Klimuk, A. Tarasau, A.B. Abacha, S.A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.T. Tran, M. Lux, C. Gurrin, O. Pelka, C.M. Friedrich, A. Garcia, S. de Herrera N. Garcia, E. Kavallieratou, C.R. del Blanco, C.C. Rodríguez, N. Vasillopoulos, K. Karampidis, J. Chamberlain, A. Clark, and A. Campello. *ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature* in Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). 2019. Lugano, Switzerland. (LNCS) Lecture Notes in Computer Science, Springer.
2. Cid, Y.D., V. Liauchuk, D. Klimuk, A. Tarasau, V. Kovalev, and H. Muller, *Overview of ImageCLEFtuberculosis 2019 - Automatic CT-based Report Generation and Tuberculosis Severity Assessment.* CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR- WS.org), 2019. ISSN 1613-0073, http://ceur-ws.org/Vol-2380/.
3. Cid, Y.D., O.A. Jiménez-del-Toro, A. Depeursinge, and H. Müller, *Efficient and fully automatic segmentation of the lungs in CT volumes.* . In: Goksel, O., et al. (eds.) Proceedings of the VISCERAL Challenge at ISBI. No. 1390 in CEUR Workshop Proceedings . No. 1390 in CEUR Workshop Proceedings, 2015.
4. Gentili, A., *ImageCLEF2018: Transfer Learning for Deep Learning with CNN for Tuberculosis Classification.* CEUR Workshop Proceedings, 2018.
5. Howard, J. et al., *fastai.* GitHub, 2018.
6. Brett, M., M. Hanke, C. Markiewicz, M.-A. Côté, P. McCarthy, and C. Cheng, *nipy/nibabel: 2.3.3* Zenodo., 2019.
7. Demsar J, C.T., et al., *Orange: Data Mining Toolbox in Python.* Journal of Machine Learning Research, 2013. **14**: p. 2349−2353.