ImageCLEF 2019: A 2D Convolutional Neural Network Approach for Severity Scoring of Lung Tuberculosis using CT Images

[1]Kavitha S [0000-0003-3439-2383], [1]Nandhinee PR, [1]Harshana S, [1]Jahnavi Srividya S and [1]Harrinei K

[1]Department of CSE, SSN College of Engineering, Kalavakkam–603110, India
kavithas@ssn.edu.in,
{nandhinee16066,harshana17053,jahnavisrividya17061,
harrinei17052}@cse.ssn.edu.in

Abstract. Tuberculosis (TB) is an air-borne disease, which affects the lungs and often spreads through sputum. According to the report of World Health Organization 9 million people world-wide are affected with TB. Tuberculosis can be cured easily when diagnosed in its early stage and with accurate CT Analysis. As an effort to form a technical forum for effective analysis and diagnosis, ImageCLEF released the Tuberculosis 2019 tasks, each dealing with one aspect of understanding and tackling the disease. We have taken up one sub-task that aims at assessing the severity of the tuberculosis disease as low or high. The task is implemented using a deep neural network approach using 2-D Convolutional Neural Network (CNN) with appropriate preprocessing. The CT volumes are segmented with the provided masks and further pre-processed with the aid of med2image, a python utility to obtain slices of CT scans, prior to training the model. The best run of the proposed CNN model resulted with an accuracy of 0.607 and an AUC of 0.626. The achieved result is placed 9th in the overall leaderboard of the ImageCLEF 2019 Tuberculosis challenge for severity scoring .

Keywords: Severity scoring; Lung tuberculosis; Pre-processing; Lung-mask; CNN; AUC; Accuracy.

1 Introduction

Tuberculosis (TB) is an airborne disease that affects the lungs. Often spread through sputum, cough and infected droplets, it is quite widespread affecting about 9 million world-wide. The treatment depends upon the degree of infection, i.e the severity [1]. The severity evaluation has been executed by medical practitioners via a diverse set of devices including mycobacterial culture test, pleural fluid and cerebrospinal fluid analysis, lesion patterns obtained from radiological images of lungs besides individualistic factors such as the patient's age, prior treatment etc. Computed Tomography (CT) is widely used for analysis of the lesion patterns. Besides being prone to errors,

a manual approach can prove to be costly, both in terms of capital and time. A computerized method, on the other hand upholds time efficiency and precision. In this paper, a Convolutional Neural Network (CNN) approach for severity scoring of lung tuberculosis based on CT scans is discussed with results. This work is a subtask of the tuberculosis tasks of ImageCLEF 2019 [2, 8]. This work establishes a standard scale against which evaluation of the CT in subject can be done for determining the severity.

The sections span across following: Section 1 gives a brief introduction about the importance of this problem and the necessity to find the severity of tuberculosis. Section 2 gives a glimpse of the dataset and how it is spread across the two classes and Section 2.1 details about the data preprocessing procedures. Section 3 explains the proposed model using convolutional neural network with the parameters chosen for analysis. In Section 4, the results of various runs are discussed. Finally, Section 5 concludes this paperwork and looks into the futuristic aspects for further improvisation of the proposed model.

2 Dataset

In this edition of ImageCLEF 2019 TB tasks, the dataset contains 335 chest CT scans of TB patients along with a set of clinically relevant metadata, where data of 218 patients are used for training and 117 for testing. For all patients, 3D CT images are stored in the compressed NIFTI (Neuroimaging Informatics Technology Initiative) file format with a slice size of 512×512 pixels and the number of slices varies from 50 to 400 for each patient. This file format stores raw voxel intensities in Hounsfield Units (HU) as well the corresponding image metadata like image dimensions, voxel size in physical units, slice thickness, etc. The selected metadata includes the following binary measures: disability, relapse, symptoms of TB, comorbidity, bacillary, drug resistance, higher education, ex-prisoner, alcoholic, smoking and severity score ranges from 1 to 5 assigned by medical doctors. To treat this task as a binary classification problem, the severity scores are grouped as high severity with scores 1, 2 and 3, and low severity with scores 4 and 5. Moreover, for all patients automatic extracted masks of the lungs are provided. In Table 1, the number of patients of each severity class in the training set and the number of patients in test set is given [2].

Table 1. Severity scoring dataset – Patient wise – Training and Test set

Severity type	Training	Testing
Low	118	117
High	100	
Total patients	218	

From the given dataset, sample images of type “high severity” class and “low severity” class is shown in Figure 1 and 2.

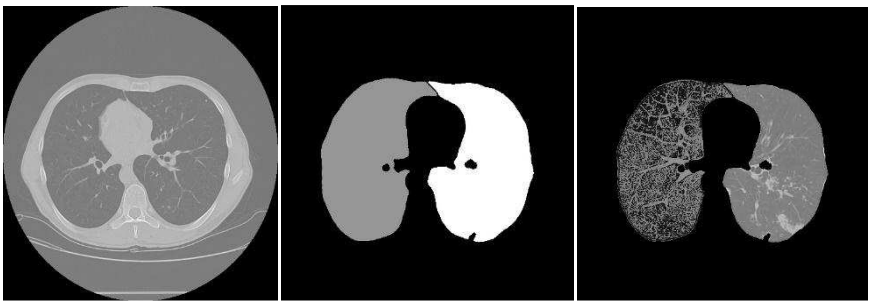

Fig. 1. "High Severity" Patient ID 196 Slice 66

Left-CT Scan of Lung from dataset, Middle-Corresponding mask of the lung, Right- Masked image.

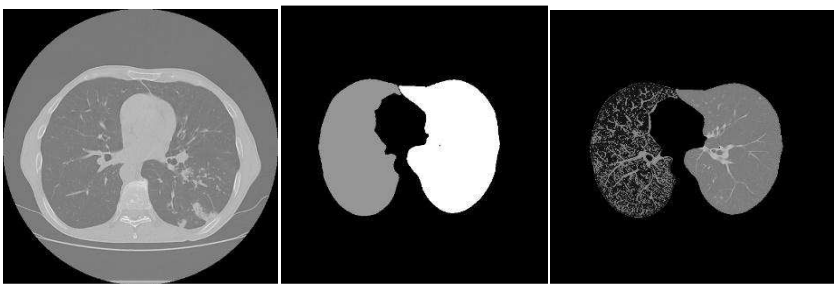

Fig. 2. "Low Severity" Patient ID 181 Slice 65

Left-CT Scan of Lung from dataset, Middle-Corresponding mask of the lung, Right- Masked image.

2.1 Data Preprocessing

The dataset for the TB tasks are given in compressed NIfTI (Neuroimaging Informatics Technology Initiative) format. Initially, the file is decompressed and the slices were extracted using med2image, a Python utility. For each Nifti image we obtain a certain number of slices ranging from 50 to 400 jpeg images. The lung masks provided by the organizers are used, to avoid potential confusion resulting from identification of similar structures resembling lungs in other parts of CT images. The next step involves masking the images. The given masks are converted to grayscale format and each pixel is checked individually; if the pixel is not black it is converted to white. In this way, a final mask is created, with pixels of two values such as black (0) or white (255). Now the original scan of the lung is converted to grayscale and each pixel of it is multiplied with the corresponding pixel in the created final mask using bitwise and operation. Thus, the lungs are segmented from the original scans [4]. On the other hand, not all slices necessarily contain relevant information that can be useful to identify severity of TB. For the same reason, it is essential to filter

slices to preserve only those that can be informative and contain relevant information. Upon visual inspection, slices ranging between 55 and 85 are used and other slices were eliminated from further processing. The slices being ordered, the 31 most informative usually fall at the center of the list. The workflow of the preprocessing stages is given in Figure 3.

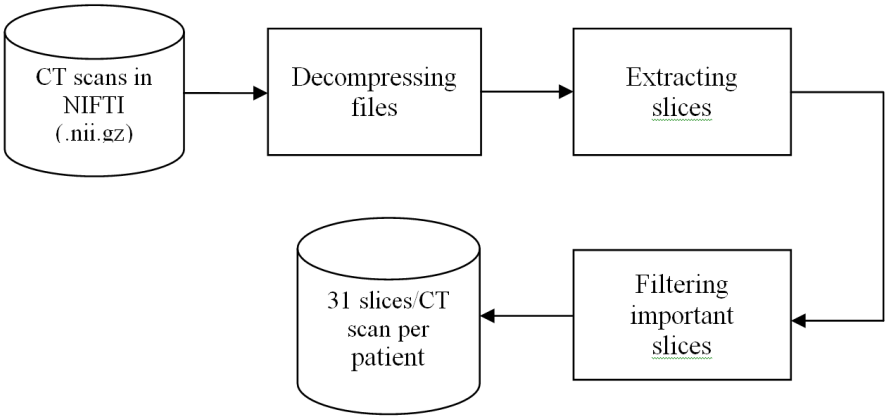

Fig. 3. General flow of the preprocessing stage

3 Methodology

Convolutional Neural network takes an image as input, passes it through a series of convolutional layers, nonlinear activation layers, pooling (downsampling) and a fully connected layer to output the classification labels. It differs from normal neural network in two aspects: atleast one convolutional layer and filters. The model for TB severity scoring is created using 2-D convolutional neural network using software libraries Keras [5] with Tensorflow [6] for backend. The 3D images of the procured CT scans are sliced and converted to 2D images in the preprocessing stage. The network is designed with three 2D convolution layers, rectified linear unit (ReLU) activation function and each convolution layer followed the max pooling layer. These led to a complete layer structure which is connected to 1000 outputs with weight by the dense layer with ReLU activation. Finally, these activations run through a softmax layer, which output a tensor of size 2, for each category. Binary cross-entropy is used as loss function with Adam and RMSProp as optimizers. The model files are built for different runs by varying the hyper parameters of the base model. The corresponding CNN design structure is shown in Figure 4. In each layer the values are mentioned from the model summary of one run for more clarity.

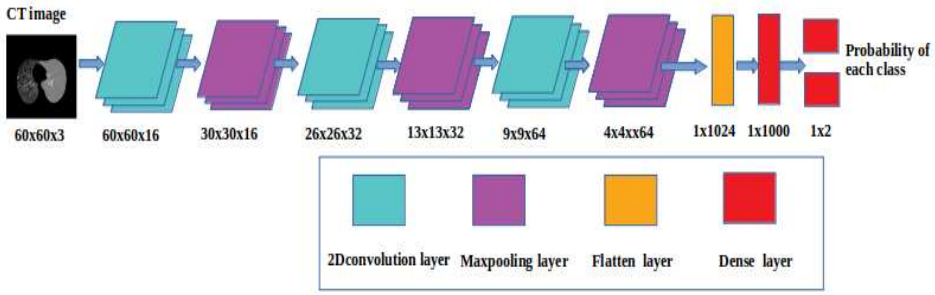

Fig. 4. Base design of the 2D CNN used for training

4 Experiments and Results

The CNN model is trained by varying the hyperparameters such as number of filters, epochs and optimizers. The runs had a filter size of 64×64 with batch size 32 and loss type as binary cross entropy. The difference in the accuracy is brought by changing the epoch value and the optimizers such as Adam and RMSProp. The different runs of CNN model by varying the hyperparameters is given in Table2.

Table 2. Different runs of the CNN model – Varying hyperparameters

Hyperparameters	Run 1	Run 2	Run 3	Run 4
No. of convolutional layers	3	3	3	3
No. of filters in each layer	16×32×64	16×32×64	**64×32×32**	**64×32×16**
Size of each filter	64×64	64×64	64×64	64×64
Pooling function	max	max	max	max
Activation functions	relu , softmax	relu, softmax	relu, softmax	relu, softmax
Batch size	32	32	32	32

Number of epochs	**15**	**20**	15	15
Loss type	binary cross entropy	binary cross entropy	binary cross entropy	binary cross entropy
Optimizer	**RMSProp**	**Adam**	RMSProp	RMSProp

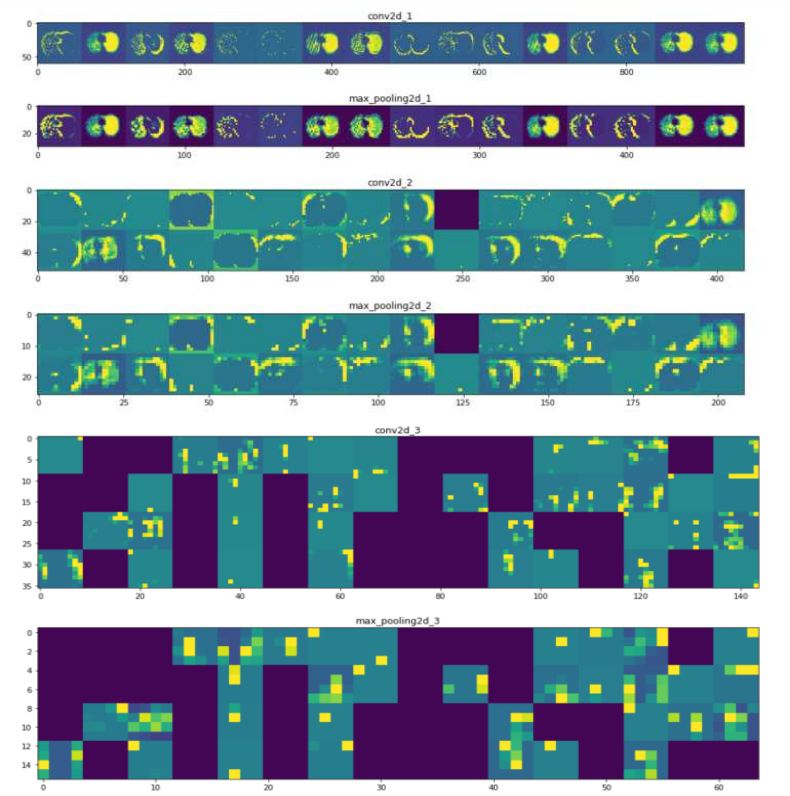

Fig. 5. Visualization of 3 convolution layers and max pooling

The intermediate visualization of convoluation layers and max pooling is shown in Figure 5, for Run1, Patient ID 181 and slice number 65.

The result of submitted four runs are listed in Table 3, for training, validation and test dataset with necessary parameters. In testing, 31 slices per patient is considered as similar to training and validation, for all 117 patients. The probability of high severity for each patient is calculated from the average of “probability of high” of all 31 slices of the specific patient. For example, the class probability of patient ID 77 in testing, for slice number 60 is represented as [0.1.]. Here, “0.” is the probability of having low severity and “1.” is the probability of having high severity. We find the

probability of having high severity for each of 31 slices of the patient and then computed the average of it. The average probability of high severity for patient ID 77 in each run is given in Table 4.

The test dataset results are evaluated using two metrics namely accuracy and Area Under ROC Curve (AUC) and ranking is carried out among the participated teams.

Table 3. Results of different runs – Training, Validation and Testing

Run No.	Training accuracy	Validation accuracy	Training loss	Validation loss	AUC	Accuracy
1	0.8314	0.8011	0.3698	0.4223	0.5446	0.5299
2	0.8491	0.8390	0.3378	0.3496	0.6067	0.5726
3	**0.8840**	**0.8434**	**0.2869**	**0.3132**	**0.6264**	**0.6068**
4	0.8754	0.8103	0.2979	0.4284	0.6133	0.5385

Table 4. Test run of Patient Id: 77 with its probability score of high severity

Test Run	Probability score of high severity
Run 1	0.67741935
Run 2	0.80645161
Run 3	0.83870968
Run 4	0.86362070

For better visualization, the same information is plotted and shown as graphs in Figures 6 and 7 using Tableau Tool. In Figure 6, the value of evaluation metrics for test set is given for all four runs. From the graph, it is clearly visible that run 3 has higher AUC and accuracy than remaining runs. In Figure 6, the accuracy for training, validation and testing dataset are given for all four runs. From the graph, it is clearly visible that run 3 has higher value in all the cases. In addition, run 4 has higher training and validation accuracy, but the test accuracy is low than run 2, might have occurred due to the chosen filter size of each layer.

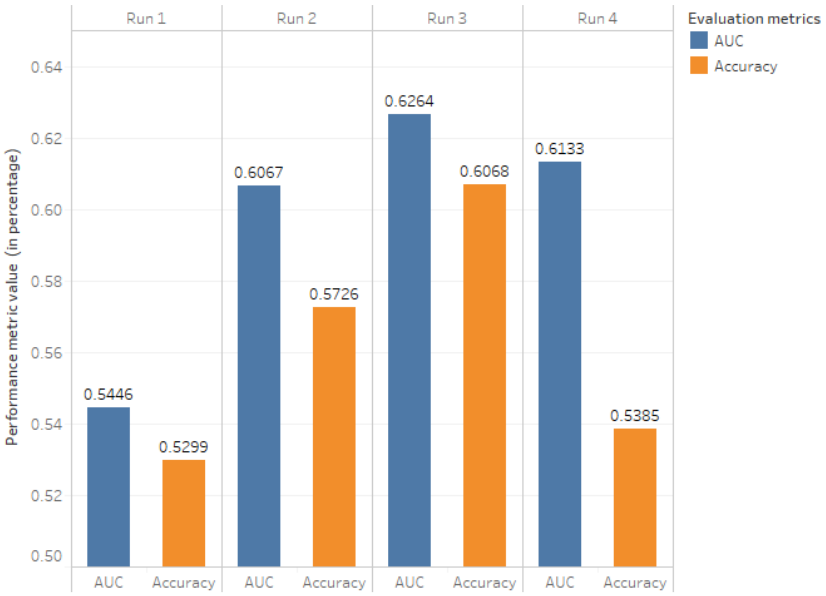

Fig. 6. Performance analysis – Runs vs metrics

Fig. 7. Comparison of accuracy between training, validation and testing

In the ImageCLEF 2019 Tuberculosis-Severity scoring subtask, 4 runs are submitted and the best run of our team is ranked 9th in overall among the teams participated is given in Table 5 [3].

Table 5. Top 10 rankings of ImageCLEF 2019 Tuberculosis - Severity scoring task

Rank	Team name	AUC	Accuracy	No. of runs submitted
1	UIIP_BioMed	0.788	0.718	2
2	SergeKo	0.775	0.718	2
3	KirillB	0.770	0.692	10
4	CompElecEngCU	0.763	0.658	2
5	agentili	0.721	0.684	9
6	yashindc(Organizer)	0.720	0.641	6
7	UniversityAlicante	0.701	0.701	10
8	MostaganemFSEI	0.651	0.615	10
9	**Kavitha**	**0.626**	**0.607**	**4**
10	Shopon	0.611	0.615	2

When the results of all runs are sorted by descending related to AUC for SVR subtask, we have obtained 29th, 31st, 35th and 43rd rank for the four runs submitted by our team [2, 7].

5 Conclusion and Future Work

In this paper, analysis of severity scoring (SVR) subtask for lung tuberculosis using 2D Convolutional Neural Network is implemented. The classification results obtained for the given set 3D CT Images are submitted for evaluation. In our approach, preprocessing of the dataset has been carried out to convert the images into 2D slices, and the images are split into training and validation set. The proposed model is built using CNN, trained and validated using tuning the hyperparameters for four different runs. From the runs submitted, the primary run is ranked 9th place among the team participations.

CNN is a preferred approach, since it facilitates automatic detection of the low level and high level features, from large training dataset. However, a large dataset might prove disadvantageous in terms of memory during the training phase. This can be overcome by the use of a GPU and choosing optimal hyperparameters. In future, the proposed model can be improvised by considering all the slices of the CT images, to build the train model using GPU and transfer learning approach.

References

1. WHO page, https://www.who.int/news-room/fact-sheets/detail/tuberculosis, last accessed May 2019
2. ImageCLEF 2019 page, https://www.imageclef.org/2019/medical/tuberculosis, last accessed May 2019
3. Crowd AI page, https://www.crowdai.org/challenges/imageclef-2019-tuberculosis-severity-scoring/ leaderboards, last accessed May 2019
4. Yashin Dicente Cid, Oscar A. Jiménez-del-Toro, Adrien Depeursinge, and Henning 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 (Apr 2015)
5. Keras documentation, https://keras.io/, last accessed May 2019
6. Tensorflow documentation, https://www.tensorflow.org/, last accessed May 2019
7. Yashin Dicente Cid, Vitali Liauchuk, Dzmitri Klimuk, Aleh Tarasau, Vassili Kovalev, Henning Müller, Overview of ImageCLEFtuberculosis 2019 - Automatic CT-based Report Generation and Tuberculosis Severity Assessment, CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380/.
8. Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos Cuevas Rodríguez, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, Antonio Campello, ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 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)
9. Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Henning Müller, Overview of ImageCLEFtuberculosis 2018 - Detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score, CLEF working notes, CEUR, 2018.