Feature and Deep Learning Based Approaches for Automatic Report Generation and Severity Scoring of Lung Tuberculosis from CT Images

Kirill Bogomasov, Daniel Braun, Andreas Burbach, Ludmila Himmelspach, and Stefan Conrad

Heinrich-Heine-Universität Düsseldorf, Institut für Informatik
Universitätsstraße 1, 40225 Düsseldorf, Germany
{bogomasov, daniel-braun,andreas.burbach,
ludmila.himmelspach,stefan.conrad}@hhu.de

Abstract. The paper presents two approaches for automatic Computed Tomography (CT) report and tuberculosis (TB) severity scoring which were two subtasks of ImageCLEFtuberculosis 2019 challenge. While our first approach uses image processing techniques for feature extraction from CT scans, our second approach uses artificial neural networks (ANN) for predicting probabilities for different lung irregularities associated with pulmonary tuberculosis and tuberculosis severity assessment. The results showed that our feature-based approach is still a competitive method that achieved rank 3 of 54 in the severity scoring subtask and rank 7 of 35 in the CT report subtask.

Keywords: automatic CT report · tuberculosis severity scoring · medical image classification · feature extraction · deep learning

1 Introduction

The tuberculosis task [5] of the ImageCLEF 2019 [10] challenge consisted of two subtasks dealing with analysis of Computed Tomography (CT) images of patients suffering from pulmonary tuberculosis. The aim of subtask #1 was the tuberculosis severity assessment based on CT scans. The subtask #2 was dedicated to the automatic generation of a CT report including the information about the left and right lung affection, presence of calcifications, presence of caverns, pleurisy, and lung capacity decrease. Both subtasks shared the same data set consisting of CT images and additional patient's meta data including information about education, imprisonment, disability, comorbidity, and others.

Last year our team participated in the severity scoring subtask at ImageCLEFtuberculosis 2018 challenge [6]. Our feature-based approach achieved rank

10 of 36 regarding the RMSE measure [2]. This result showed that our methods could compete with more complicated and computationally intensive methods in the field of deep learning. Since our feature-based approach provided a descriptive image classification framework, we decided to improve and to adapt it to the requirements of both subtasks of the ImageCLEF 2019 challenge [5]. On the other hand, taking the last years research trends into account, we developed a new deep learning-based approach.

2 Feature Based Approach for Automatic CT Report Generation and Tuberculosis Severity Scoring

In this section we describe our feature-based approach for automatic CT report and severity score prediction from CT scans. The main motive for developing a feature-based approach was the ability not only to predict the probabilities for different lung irregularities but also the ability to mark them in CT scans. This could also be helpful for physicians during manual assessment of CT scans. Furthermore, our approach provides information about the influence of different lung damages and additional patient's data on the tuberculosis severity score.

2.1 Preprocessing

Some features that we used for the automatic CT report were extracted from the original CT scans, while other features were easier to extract from binary images. Therefore, we binarized all CT scans using IsoData method [13]. We used lung masks for extraction of all features for the CT report task. Some of the lung masks that were provided by the organizers of the task [7] still did not cover large lesions. For this reason we decide to use our own lung masks extracted by the segmentation algorithm described in [4]. This algorithm examines the silhouettes of extracted masks for irregularities and reconstructs the masks. Although the reconstructed lung masks did not perfectly cover the entire lung, they still contained more lung pixels than lung masks provided by the organizers of the task.

2.2 Automatic CT Report Generation

Presence of Calcification Pulmonary calcification in CT scans was determined for left and right lung separately depending on the number of pixels that were identified as part of calcification. Since different Hounsfield Unit (HU) ranges for pulmonary calcification in CT scans were proposed in the literature [3, 8, 12] and the Hounsfield Units were not standardized in CT scans in the data set, we decided on a relatively large range between 300 HU and 3000 HU. In this way, we were able to identify calcifications of different density. On the other hand, our range for calcification contains the HU range for bones that were often erroneously covered by the lung masks. To reduce the presence of bones in the examined lung area, we adjusted the lung masks in a preprocessing

step by removing pixels of their boundaries along the z-axis using morphological erosion function [11] with a disk of radius four pixels. Since many CT scans contained noise patches that could be erroneously classified as calcified nodules, we removed all objects smaller than 10 pixels that were identified as calcifications. Finally, we added up the pixels of found calcifications over all CT scan slices along the z-axis in the file. If either left or right lung or both contained more than 400 calcification pixels, we stated the probability of presence of lung calcifications as 1 otherwise as 0. This threshold value was determined based on the cross-validation Area Under the ROC Curve (AUC) value for presence of calcification on the training set.

Since Hounsfield Unit range for plastic and metal overlaps our range for calcification, our method for detection of calcification presence tended to false positives for patients that had medical appliances in the lung. To prevent misclassifications in such cases, the shape of found calcifications could be additionally examined.

Presence of Caverns At ImageCLEFtuberculosis 2018 [6], we used a simple approach for detection of pulmonary caverns. The principal idea of the method was detecting caverns as dark spots surrounded by light tissue in binarized CT image slices along the z-axis [2]. The main weak point of our approach was that trachea and bronchi were incorrectly recognized as caverns. Therefore, we cut out the middle part of the lung to avoid false positives. Unfortunately, that workaround has led to many false negatives because our method did not detect caverns that were either completely or partly located in the cut out part of the lung. For this reason we improved our last year approach for detection of pulmonary caverns by examining the entire lung.

The Fleischner glossary defines pulmonary cavities as thick-walled gas-filled spaces [9]. The main difference to trachea and bronchi is that cavities are completely covered by cavity walls. Therefore, we validated a cavern in a binarized CT scan slice along the z-axis as such only if its pixels were detected as pixels of a cavern in the CT scan slices along the x- and y-axes. We estimated the volumes of pulmonary caverns and their walls for right and left lung separately by adding up the pixels of validated cavities and cavity walls over all CT image slices along the z-axis. We used these four features for training a linear regression model for predicting the presence of caverns.

Our improved method reliably detected caverns in CT scans in the training set as long as the distances between the slices in the scans were not too large so that all cavity walls were depicted in the CT images. Unfortunately, our approach still produced false positives due to artifacts on the images mainly caused by the heartbeat of patients. Therefore, an additional preprocessing step is needed for elimination of artifacts in CT scans.

Presence of Pleurisy Pleurisy is inflammation of pleura which is a thin membrane that covers the lungs [1]. Since inflammation often leads to thickening of the tissue and pleura thickening increases the distance between the lung and

bones, in our approach for pleurisy detection, we compared the average distance between the boundaries of the lung masks and bones in images along the z-axis in patients with and without pleurisy. For that purpose we overlayed the lung masks and the bone masks which represent pixels of the original CT scan with Hounsfield Units between 300 and 3000. In the resulting image, we calculated the average distance between pixels of the lung mask boundaries and the nearest bone pixels. Then we averaged the distances between lung and bones over all CT scan slices along the z-axis for right and left lung separately and used them for training a linear regression model for pleurisy prediction.

Lung Capacity Decrease The lung capacity is the maximum amount of air that the lung can hold. Some kinds of lung tissue damage caused by Mycobacterium tuberculosis (MTB) bacteria may decrease the capacity of the lung. Since an automatic detection and classification of different types of lung lesions from CT scans is a challenging problem, we predicted the probability of the lung capacity decrease based on the estimated ratio of the lung tissue to the entire lung volume. Assuming that the lung tissue ratio compared to the lung volume is larger in patients with decreased lung capacity than in patients with normal lung capacity, in our approach, we did not differentiated between healthy and damaged lung tissue. Similar to our last year approach [2], we calculated the ratio of the lung tissue as a relation of white pixels in the binarized CT image to the number of pixels in the lung mask averaged over all slices along the z-axis. Finally, we trained a linear regression model for lung capacity decrease prediction using the ratios of the lung tissue for left and right lungs as features.

Right and Left Lungs Affected Mycobacterium tuberculosis (MTB) bacteria causes more kinds of lung damage than calcifications, caverns, pleurisy, and lung capacity decrease. Therefore, the estimation model for probability of lung affection based on the probabilities for lung damage described before did not achieve satisfactory results on the training set. On the other hand, raw feature values that we extracted for predicting the probability of aforementioned lung damage, provided more information about further lesions in the lung. For this reason, we used the number of calcification pixels in the lung, average distance between the lung and bones, and the ratio of lung tissue to the lung volume for left and right lung, separately, as features for training random forests models for predicting the probabilities of affection of lungs.

2.3 Tuberculosis Severity Scoring

At ImageCLEFtuberculosis 2018 [6], our system achieved its best results for tuberculosis severity score prediction using three features: the cavern volume, the volume of cavern walls, and the infection ratio [2]. This year we used data from the CT report task combined with provided patient's meta data. Using linear regression as classifier, we obtained the 5-fold cross-validation AUC of approximately 0.8 for severity score on the training set. The most important features

for severity score prediction were the probability of left and right lung affection, information about the imprisonment, the probability for pleurisy, and information about education. Although some features seemed to play an insignificant role, their elimination diminished the AUC value for severity score. Since some features from meta data were very important for severity score prediction, we tested our linear regression model on the training set only using patient's meta data. We obtained AUC value of approximately 0.75. On the other hand, the linear regression model trained only using data from CT report achieved the same AUC value. Although we were aware that feature values predicted for the CT report task were inaccurate to some degree, we used them combined with provided patient's meta data for training a linear regression model for TB severity score prediction.

3 Deep Learning Based Approach

Deep Learning has been applied on solving medical relevant research questions. Among other things it is used for classification of brain and lung tumors. Thus, Liu and Kang [17] for example achieves an AUC value of 0.981 with their ANN on the LIDC-IDRI data set [18] for the binary classification of lung cancer.

In addition to the classification of the CT scans into the predefined disease stages, the task can be subdivided into a further subtask, namely the segmentation. We suspect that the occurrence of disease-typical symptoms, such as calcification, caverns and pleurisy, may help in the subsequent classification. The topic of the localization and classification of objects is the subject of many scientific publications.

Some of the most promising approaches are based on the U-Net architecture [15]. This is shown, for example, by the fact that the winner of the 2018 BraTS Challenge used a U-Net variant [16]. The BraTS data set contains of CT scans of brain tumor patients and is therefor similar to the given tuberculosis data.On the one hand an advantage of the U-Net architecture is that the network considers the semantic context of the entire image during segmentation, on the other the hand U-Net architecture needs only a small amount of training examples to produce good results. Regarding the low amount of training data of the two tuberculosis tasks, this is a sufficiently important feature. We will use one architecture for both tasks, severity scoring and CT report, with the only difference being the number of final classifications to represent the different amount of possible labels. Isensee et al. showed that the architecture of the U-Nets is already so high-performant that a meaningful pre- and post-processing offers a greater potential for improvement than the change of the architecture [14]. Therefore, we start our processing pipeline with preprocessing and extend the architecture of the original U-Net [15] by an additional classification CNN. Afterwards we finish our approach with postprocessing. The exact explanation follows in the next sub-chapters.

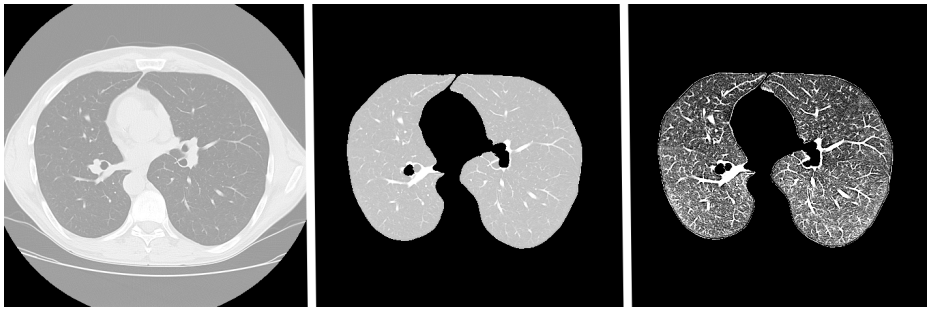

Fig. 1. Left: No preprocessing. Middle: Only segmentation. Right: Full preprocessing.

3.1 Preprocessing

The data set contains several anomalies which make preprocessing necessary. The CT scans in the given data set have 3 different values $\{-3024, -2048, -1024\}$ for "outside of body" - mark. Probably because the images were taken by different scanners and are not standardized. For this reason, some serious jumps can be found in the value ranges of the Hounsfield Units. Beside of that, there are even higher values for some noisy pixels. Similar to [19], we used a four-stage preprocessing to standardize the CT scans.

- Step 1: Remove empty gap. "NULL"-representing pixel values outside of body are often much lower than the values inside. To prevent that no area of the examination remains empty, each "NULL" - representing pixel is replaced with the next higher value.
- Step 2: Removing noise by range limits. The new value range is limited to [-1000; +2000]. Outside pixel values are set to the limit value.
- Step 3: Min-max normalization to [0,1].
- Optional Step 4: In the following the lung area is segmented with the binary masks from the original data set 1. Finally we reduced the image size by removing "0"-values in border area.

Figure 1 shows the three options of preprocessing.

3.2 Architecture

As mentioned previously, our chosen architecture is based on the original U-Net approach, but we changed the original 2D convolutional layer to 3D. Additionally we added a final classification CNN, based on the well known VGG19 architecture [20], for a binary output, since we have a two-classes problem. Figure 2 shows a draft of the resulting network architecture.

During the training and in the later classification we limit the input to 16-slice sliding windows, which contains coherent slices along z-axis, for two reasons. First, this reduces the requirements on GPU memory. Second, we now have a fixed input depth without the need to scale it. This not only serves to reduce

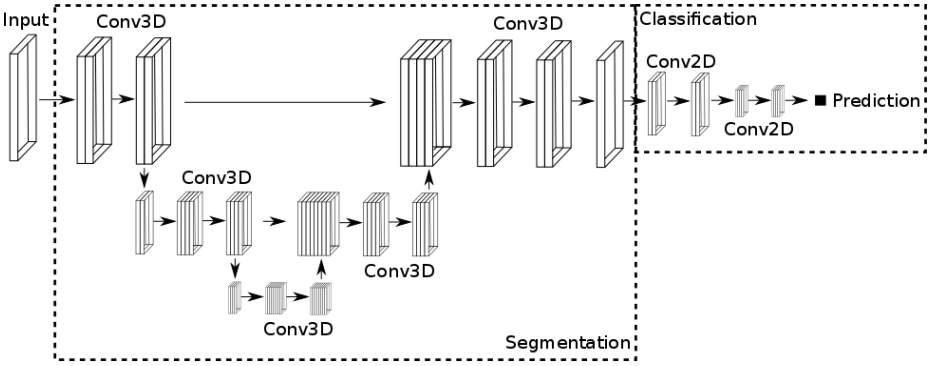

Fig. 2. The architecture of the proposed network.

the requirements on GPU memory, but has also proven to be a useful value to enhance the precision. Complementary, a more accurate prediction is produced, because of a several classification results for each image. In the next step, we halve the image, separating it into left and right lungs. This distinction is not taken into account during training. Finally, we scale the input data to 192×256 with a bilinear interpolation. This results in an input tensor of $192 \times 256 \times 16$.

For the U-Net, as segmentation network, we chose a depth of four with a number of eight filters for the first convolutional layers. We use maxpooling for the downscaling path and a transposed convolution for the upscaling path. Furthermore, batch normalization is applied after each convolution and a dropout value of 30% for the last convolutional layer in the downscaling path. As activation function we use the rectified linear unit. To get our segmentation mask we use a convolutional layer with filter size one in each direction. For task one this results in one segmentation mask, due to the fact that we have a binary classification. Contrary to that, we use five segmentation masks for task two. Even though there exist six labels in the task, we only need one probability to distinguish the affection of the left or right lung due to the splitting of the lung as preprocessing. An example of different segmentation masks for task two can be seen in Figure 3.

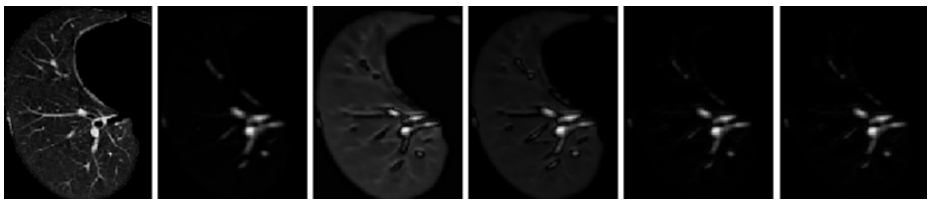

Fig. 3. The left image shows the input slice. All others show the activations in the different segmentation masks.

Algorithm 1 Definition of Max-Rule

Require: $\tau \in \mathbb{N}$
  if $|D_{left} - D_{right}| > \tau$ **then**
    $P \Leftarrow max(S_{pos} \cup S_{neg})$
  else
    if $|S_{pos}| = |S_{neg}|$ **then**
      if $1 - max(S_{pos}) < min(S_{neg})$ **then**
        $P \Leftarrow max(S_{pos})$
      else
        $P \Leftarrow min(S_{neg})$
      end if
    else
      if $|S_{pos}| > |S_{neg}|$ **then**
        $P \Leftarrow max(S_{pos})$
      else
        $P \Leftarrow min(S_{neg})$
      end if
    end if
  end if
  return P

The segmentation mask is used as input in our final classification CNN. For the CNN we also use a depth of four with eight as number of filters for the first convolutional layers. Like for the segmentation network, batch normalisation for all and a 30% dropout for the last convolutional layer are applied. A leaky rectified linear unit is used as activation. The final layer is a dense layer with one neuron to represent the probability of the label. In task one we have one classification network, but for task two we use five independent classification networks, one for each label.

3.3 Postprocessing

The network predicts the class of 16-slice windows of the CT scan. To get an overall prediction P for a whole CT-scan, an aggregation of a set of predictions has to be made. Therefore we divide each CT scan into three sections of same size. For each of these sections a prediction p_i with $\{p_i \in \mathbb{R} | 0 \leq p_i \leq 1 \ \wedge i \in \{1, \ldots, 6\}\}$ is calculated. Taking into account the left and right half, we get a total of six results.

Now we propose four methods to merge these six partial results p_i into one final result P.

1. Average: The result is defined as $P = \overline{\{p_i | i \in \{1, \ldots, 6\}\}}$, namely the average prediction value over all partial predictions.
2. Max-Rule: For this rule we define D_{left} and D_{right} as the number of lung slices in z-direction of the left respectively right lung. Also let S_{pos} be the set of positive predictions for which holds that $p_i \geq 0.5$ with $i \in \{1, \ldots, 6\}$.

Similary, S_{neg} is the set of negative predictions defined as $S_{neg} = \{p_i < 0.5 | i \in \{1, \ldots, 6\}\}$. Like Algorithm 1 shows, we first check if part of the lungs is missing. This occurs due to the fact that the size of the left and the right lung can diverge due to the preprocessing while reducing the zero values at the image borders. Consequently, we make the assumption that this difference is a sign of serious illness. Therefore, if the difference between D_{left} and D_{right} exceeds a threshold $\tau \in \mathbb{N}$, the maximal partial prediction value p_i is chosen as probability. If the depth of the lung does not differ too much, we let the majority decide and therefore choose the maximum respectively the minimum value from the set, S_{pos} or S_{neg}, that has more elements. If the two sets have an equal amount of elements, the value with the smallest distance to the respective target value 0 or 1 is chosen.

3. Average-Rule: Similar to Max-Rule, the only difference is that the calculation of the resulting prediction value P does not select the maximum or minimum but the average over all values of the corresponding result set S_{pos} respectively S_{neg}.
4. Confidence correction: For each window of a CT scan from the validation data set, consisting of 16 slices, the coefficient which is necessary to change the prediction of the respective window, is calculated so that the classification result is the correct class.

4 Evaluation and Results

This section shows final performance results of submitted runs in the severity scoring (subtask #1) and CT report (subtask #2) challenge. The final ranking in the severity task was done based on the Area Under the ROC Curve (AUC) value, while the final ranking in the CT report task was done based on the average AUC value. Table 1 summarizes the results for Top-10 submitted runs with the highest AUC value and the best run for our deep learning-based approach for

Table 1. Short overview of submitted runs for subtask 1 – Severity scoring.

Group name	Run	AUC	Accuracy	Rank
UIIP_BioMed	SRV_run1_linear.txt	0.7877	0.7179	1
UIIP	subm_SVR_Severity	0.7754	0.7179	2
HHU	**SVR_HHU_DBS2_run01.txt**	**0.7695**	**0.6923**	**3**
HHU	SVR_HHU_DBS2_run02.txt	0.7660	0.6838	4
UIIP_BioMed	SRV_run2_less_features.txt	0.7636	0.7350	5
CompElecEngCU	SVR_mlp-text.txt	0.7629	0.6581	6
San Diego VA HCS/UCSD	SVR_From_Meta_Report1c.csv	0.7214	0.6838	7
San Diego VA HCS/UCSD	SVR_From_Meta_Report1c.csv	0.7214	0.6838	8
MedGIFT	SVR_SVM.txt	0.7196	0.6410	9
San Diego VA HCS/UCSD	SVR_Meta_Ensemble.txt	0.7123	0.6667	10
...	...	...	...	...
HHU	run_6.csv	0.6393	0.5812	27

Table 2. Short overview of submitted runs for subtask 2 – CT report.

Group name	Run	Mean AUC	Min AUC	Rank
UIIP_BioMed	CTR_run3_pleurisy_as_SegmDiff.txt	0.7968	0.6860	1
UIIP_BioMed	CTR_run2_2binary.txt	0.7953	0.6766	2
UIIP_BioMed	CTR_run1_multilabel.txt	0.7812	0.6766	**3**
CompElecEngCU	CTRcnn.txt	0.7066	0.5739	4
MedGIFT	CTR_SVM.txt	0.6795	0.5626	5
San Diego VA HCS/UCSD	CTR_Cor_32_montage.txt	0.6631	0.5541	6
HHU	**CTR_HHU_DBS2_run01.txt**	**0.6591**	**0.5159**	**7**
HHU	CTR_HHU_DBS2_run02.txt	0.6560	0.5159	8
San Diego VA HCS/UCSD	CTR_ReportsubmissionEnsemble2.csv	0.6532	0.5904	9
UIIP	subm_CT_Report	0.6464	0.4099	10
...	...	...	...	...
HHU	CTR_run_1.csv	0.6315	0.5161	12

severity scoring task. Table 2 lists the results for Top-10 submitted runs with the highest mean AUC value and the best run for our deep learning-based approach for CT report task. In the following subsection we describe the results for our approaches in detail.

4.1 Evaluation Results for the Feature Based Approach

Since we used results from the CT report task for TB severity score prediction, it is more sensible to start describing results for the CT report task. As highlighted in Table 2, our best run for the feature based approach was ranked on the seventh place. In this run we predicted the probabilities for lung irregularities as described in Section 2.2. In our second best run, we predicted the probability of presence of caverns only based on the number of cavern pixels in left and right lungs, separately, omitting the pixels of cavern walls. This run was ranked on the eighth place which is a worse result. Unfortunately, we did not receive the detailed evaluation results, so we can not comment on the performance of our approach regarding prediction of other lung irregularities.

In severity scoring task, the best run for our feature based approach was ranked on the third place among 54 submitted runs. In this run we predicted the severity score using patient's meta data and the results from our best run in CT report task. The prediction of severity score in our second best run was based on patient's meta data and the results from our second best run in CT report task. Although we did not submit a run for TB severity score predicted only on the basis of provided patient's meta data, the results for these two runs showed a positive impact of results from the CT report task on the tuberculosis severity score prediction.

Table 3. Deep Learning-based Approach for Severity Scoring.

Run name	AUC	Accuracy	Preprocessing	Postprocessing	Data
run_06	**0.6393**	0.5812	-	method 1	validation split
run_08 [1]	0.6258	**0.6068**	mixed	method 1	validation split
run_04	0.6070	0.5641	complete	method 1	validation split
run_07	0.6050	0.5556	complete	method 3, $\tau = 5$	all data
run_03	0.5692	0.5385	complete	method 3, $\tau = 10$	validation split
run_05	0.5419	0.5470	segmentation only	method 1	all data
baseline	0.5103	0.4872	complete	method 2, $\tau = 5$	validation split
run_02	0.4452	0.4530	complete	method 4	validation split

4.2 Evaluation Results for the Deep Learning Based Approach

For our evaluation we used different input data. We differentiated between train-/validation split and the complete dataset as training basis. The validation set consists of 10 images.

For Severity Score Task we set up the preprocessing, as shown in Table 3. For our runs, we used either full preprocessing, just segmentation or no preprocessing at all. *Run_*08 is an exception, therefore we took an average of *run_*5, *run_*6 and *run_*7. Table 3 shows the list of postprocessing configurations of each run.

The highest AUC score is achieved by *run_*06. In this case the network got the raw input data. We presume that the good AUC score is due to the fact that the network finds relevant points outside our region of interest, which is removed through preprocessing. This can be supported by the fact that the segmentation alone generates the worst results. However, the accuracy of *run_*06 is lower than that of *run_*08. It is interesting that no neural network from those three, that we calculate the average on, can achieve such a high accuracy by itself. It seems that the networks found different features and learned differently, so in the connection they complemented each other and the accuracy increased. Surprisingly, with an accuracy of 0.453, *run_*02 score performed significantly worse than the other constellations. Presumably, this is because of our validation set size of only 10 images, which is potentially too small. And thus, the calculated coefficients cannot be generalized.

Since we had only a limited amount of runs for CT reportings, we decided to use only those constellations, that were trained on the whole data set. Because it seemed to be more reasonable to train on more data. Table 4 shows the results. The greatest value for Mean AUC of 0.6315 and Min AUC share *CTR_run_*1 and *CTR_run_*2. Compared to the third run, this shows, that for this task the preprocessing may be more valuable as for task 1. *CTR_run_*3.*csv* shows rather moderate results of 0.561 Mean AUC, which is still better than random, but still leaves space for improvement.

[1] conglomerate of run_5, run_6 and run_7

Table 4. Deep Learning-based Approach for CT Report.

Run name	Mean AUC	Min AUC	Preprocessing	Postprocessing	Data
CTR_run_1.csv	0.6315	0.5161	complete	method 1	all data
CTR_run_2.csv	0.6315	0.5161	complete	method 1	all data
CTR_run_3.csv	0.5610	0.4477	segmentation only	method 1	all data

5 Conclusion

In this paper we have shown that our feature-based approach is still competitive to our deep learning-based method and to methods of other participants of the tuberculosis task. Our best run achieved the third place regarding the AUC value in the severity assessment subtask and the seventh place regarding the mean AUC value in the CT report subtask. Although the results obtained by our approach are promising, we still see potential for improvement of our approach to achieve even better results in both subtasks.

Regarding that our neural network was not as deep as other networks in the literature, our results are promising. Especially the U-Net architecture seems to be beneficial and can be a good starting point for more research. Our preprocessing was only beneficial for subtask #2, which is surprising and therefore it would be interesting to investigate which parts of the lung had an effect on the resulting predictions. Data augmentation unexpectedly led to bad results in our first tests and we therefore refrained from using it. But we like to further investigate the usefulness of data augmentation for this task in combination with our network. Furthermore, we will test the network on other data sets, especially with segmentation data to train the U-Net separately. We hope that by this the segmentation layers will find meaningful areas, that can show us symptoms of such diseases. And regarding the results for subtask #2, more training epochs would be surely beneficial too and therefore the training will continue.

References

1. Berger, H.W., Mejia, E.: Tuberculous Pleurisy. Chest **63**(1), 88 – 92 (1973)
2. Bogomasov, K., Himmelspach, L., Klassen, G., Tatusch, M., Conrad, S.: Feature-Based Approach for Severity Scoring of Lung Tuberculosis from CT Images. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum (2018)
3. Brooks, R.A.: A Quantitative Theory of the Hounsfield Unit and Its Application to Dual Energy Scanning. Journal of Computer Assisted Tomography **1**(4), 487–493 (1977)
4. Burbach, A.: Automatic Lung Extraction from CT Scans. Bachelor's Thesis (2018)
5. Dicente Cid, Y., Liauchuk, V., Klimuk, D., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2019 - Automatic CT-based Report Generation and Tuberculosis Severity Assessment. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, ISSN 1613-0073, <http://ceur-ws.org/Vol-2380/>, Lugano, Switzerland (September 9-12 2019)

6. Dicente Cid, Y., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEF-tuberculosis 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 2018)
7. Dicente Cid, Y., Jiménez del Toro, O.A., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in ct volumes. In: Proceedings of the VISCERAL Anatomy Grand Challenge at the 2015 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 31–35. CEUR-WS (2015)
8. Grewal, R.G., Austin, J.H.M.: CT Demonstration of Calcification in Carcinoma of the Lung. Journal of Computer Assisted Tomography **18**(6), 867–871 (1994)
9. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Mller, N.L., Remy, J.: Fleischner Society: Glossary of Terms for Thoracic Imaging. Radiology **246**(3), 697–722 (2008)
10. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
11. Jankowski, M.: Erosion, dilation and related operators. In: Proceedings of 8th International Mathematica Symposium (2006)
12. Khan, A.N., Al-Jahdali, H.H., Allen, C.M., Irion, K.L., Al Ghanem, S., Koteyar, S.S.: The calcified lung nodule: What does it mean? Annals of Thoracic Medicine **5**(2), 67–79 (2010)
13. Ridler, T., Calvard, S.: Picture Thresholding Using an Iterative Selection Method. IEEE Transactions on Systems, Man and Cybernetics **8**(8), 630–632 (1978)
14. Isensee, F. et al.: No New-Net. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 234-244. Springer International Publishing (2019).
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical image computing and computer-assisted intervention. pp. 234-241. Springer (2015)
16. Isensee, F. et al.: Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: International MICCAI Brainlesion Workshop. pp. 287-297. Springer (2017)
17. Liu, K., Kang, G.: Multiview convolutional neural networks for lung nodule classification. In: Int. J. Imaging Syst. Technol, vol. 27, pp. 12-22. Wiley (2017). https://doi.org/10.1002/ima.22206
18. Armato III et al.: Data From LIDC-IDRI. The Cancer Imaging Archive. 2015
19. Braun, D., Singhof, M., Tatusch, M., Conrad, S.: Convolutional Neural Networks for Multidrug-resistant and Drug-sensitive Tuberculosis Distinction. In: CLEF2017 Working Notes, CEUR Workshop Proceedings, Dublin, Ireland. CEUR-WS (2017)
20. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: arXiv 1409.1556. arXiv preprint (2014)