ImageCLEFmed Tuberculosis 2019: Predicting CT Scans Severity Scores using Stage-Wise Boosting in Low-Resource Environments

Augustus Tabarcea, Valentin Rosca, and Adrian Iftene

"Alexandru Ioan Cuza" University, Iasi, Romania
{augustus.tabarcea,valentin.rosca,adiftene}@info.uaic.ro

Abstract. Tuberculosis (TB) still remains in our days a persistent threat and a leading cause of death worldwide. The different types of TB require different treatments, usually with antibiotics, and therefore the detection of the TB type and the evaluation of the severity stage are very important. In the ImageCLEF 2019 Tuberculosis, our group submitted a solution that addresses the problem of tuberculosis' severity prediction in low-resource environments by attempting to minimize the information required from the CT scan using a regularized variant of the SAMME.R algorithm.

Keywords: Tuberculosis · Boosting · ImageCLEF

1 Introduction

Tuberculosis (TB) is a disease caused by bacteria (Mycobacterium tuberculosis) that most often affects the lungs. Even though it is curable and preventable, tuberculosis is the second biggest killer, globally (after HIV infection) [15]. In 2016, an epidemiological research estimated 10.4 million new cases and 1.7 million deaths. Tuberculosis is spread from person to person through the air. When people with active infection cough, sneeze or spit, they propel the TB germs into the air. A person needs to inhale only a few of these germs to become infected. About one-quarter of the world's population has latent TB, which means people have been infected by TB bacteria but are not (yet) ill with the disease and cannot transmit it. People infected with TB bacteria have a 5-15% lifetime risk of falling ill with TB [17].

The 2019 edition of ImageCLEF (the Image Retrieval and Analysis Evaluation Campaign of the Cross Language-Evaluation Forum) [11] targets two tasks in the direction of serving the tuberculosis diagnosis [4]. The first task involves automated detection of tuberculosis severity, while the second one involves a computed tomography report based on CT scans and the patient's metadata.

The CT report task is a binary multi-class classification problem that addresses the presence of calcification, presence of caverns, presence of TB in the right lung, presence of TB in the left lung, pleurisy and lung capacity decrease. The Severity scoring task is a binary classification problem centered around labeling the TB severity as being "HIGH" or "LOW". Each instance of training data is labelled into one of 5 classes, where 1, 2 and 3 represents "HIGH" severity and labels 4 and 5 represents "LOW" severity.

The dataset provided by ImageCLEFmed Tuberculosis is composed of 218 patients CT scans for training and 117 for the test, each with a collection of metadata regarding the respective patient. The metadata is supplied, for better classification, in consideration to the elaborated process a doctor conducts to give a diagnostic. Each CT scan is made of slices (Section 1, that number varies from 50 to 400 and every slice has a dimension of 512 x 512.

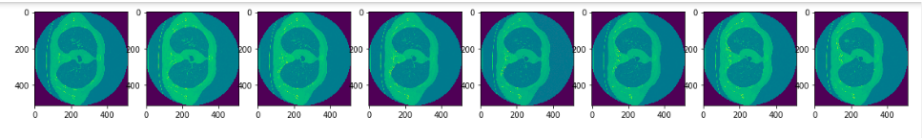


Fig. 1: Slices of lung example

Considering the need for a rapid and reliable TB treatment regimens, that is also cost effective, we investigate for the best learning algorithm that qualifies our requirements and outputs the patient condition's severity.

Our solution addresses the problem of tuberculosis severity prediction in low-resource environments by attempting to minimize the information required from the CT scan. This approach is benefic for low-buget medical organizations and small practices as they may not have access to high-end computational resources needed for a computer aided diagnosis.

2 State of the art

A number of different approaches were proposed for the task of severity prediction. More notably, in the second edition of ImageCLEFmed TB 2018 [5][12], the team UIIP_BioMed [14] managed to obtain the highest kappa score, 0.2312, with an accuracy of 0.4227 and root mean squared error of 0.7840. Furthermore, the team MedGift [6] scored 0.7708 in terms of ROC AUC, the highest in that year.

The methods used by the two teams consisted of a Deep Convolutional Neural Networks model submitted by UIIP_BioMed and an SVM with RBF kernel submitted by MedGift team. It is suitable to state that each of the two teams had completely different pre-processing procedures.

Correlation of better outcomes given more clinical data and also the CT scans of patients were studied before in [13], as a result in 2019 more metadata about the patients were given. In the edition of 2019, the best Accuracy was given also by UIIP_BioMed, with a score of 0.7350.

Our experience in prediction algorithms comes from prediction of cryptocurrency market [3] and in predicting of user activities using GPS Collections [1]. In the same time, we experienced with image processing in medical domain [7], in photo annotation task of CLEF2010 [10], and in combining semantic resources for image retrieval [9], [19].

3 Architecture

3.1 Data pre-processing

The CT scans were preprocessed trough methods whose foundation relies upon [21]. Depending on the procedure and device that was used to make the CT scans, the slices' pixel spacing vary and, as a result, their number diverge for each patient.

To simplify the process of learning we resampled the images slices number to their mean across patients, that implying approximately 130 vertical slices. Because we were limited by the available resources we decreased the size of each slice from 512 x 512 to 256 x 256. Each pixel is a signal on the Hounsfield Unit (HU) scale (Table 1), hence, we memorize each pixel as a float16 type to spare memory and aid in accelerated learning (Figure 2).

Table 1: Hounsfield units scale

Substance	**HU**
Air	-1000
Lung	-500
Fat	-100 to -50
Water	0
CSF	15
Kidney	30
Blood	+30 to +45
Muscle	+10 to +40
Grey matter	+37 to +45
White matter	+20 to +30
Liver	+40 to +60
Soft Tissue, Contrast	+100 to +300
Bone	+700 to +3000

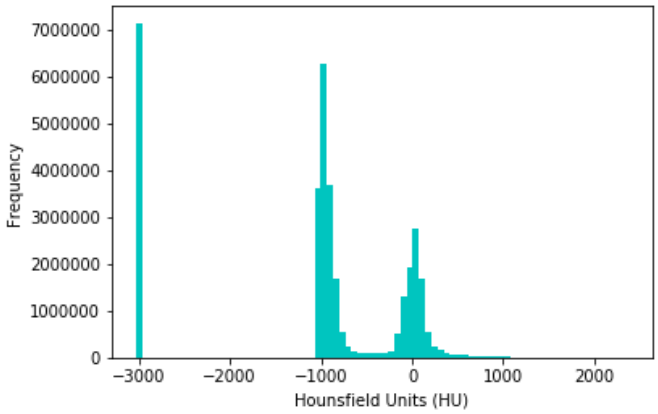
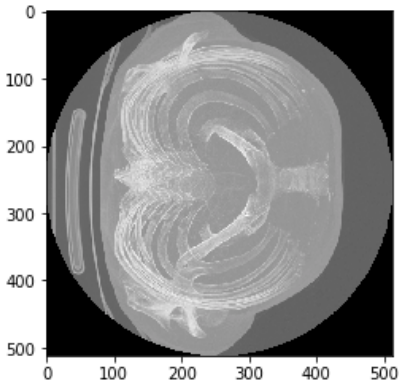


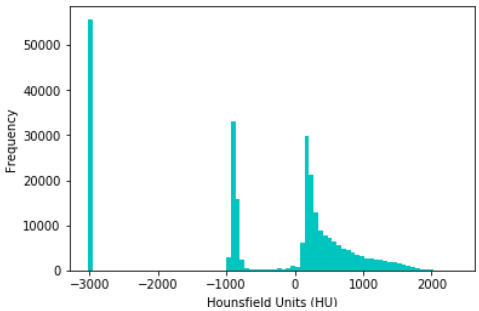
Fig. 2: HU pixels value on a patient

The HU signals for bones can reach up to 3000+ units, starting from around 400 units. Figure 3a shows merged slices of a patient for a more reliable overview of the patient bones section (HU signal values representing bones on axial view were plotted in Figure 3b).

The pixels representing bones, whose value is greater than 400, were replaced by those representing air for a better focus on the tissue of the lungs, both affected and healthy, so any classifier could distinguish among those two categories. Afterward, we normalized the values between [0, 1].



(a) Axial slice overview of bones structure



(b) Maximal HU pixel value in each axial point

Fig. 3: Axial slice overview (left) and Maximal HU pixel value (right)

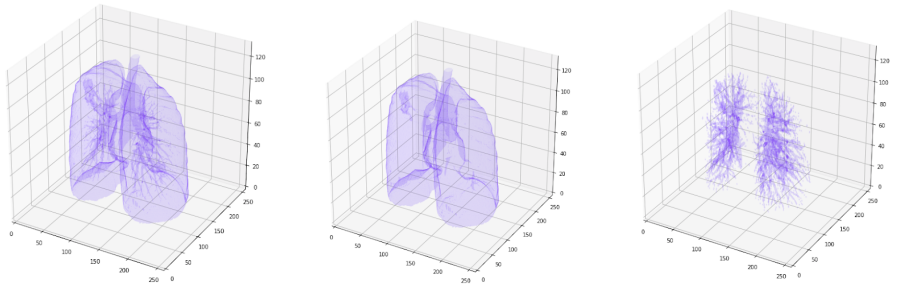
To summarize the pre-processing of our data, these steps were applied:

- resample the height of the lung voxel (number of slices per patient);
- resample the width and height of the lung voxel (all slices of each patient);
- converting the signals to float16 type (by doing so no information is lost);
- replace pixels representing bones structure from the lung voxel;
- normalize the pixel values.

3.2 Further propossed pre-processing methods

Further pre-processing methods were conceptualized although they weren't exploited because of a lack of time and scarce computational resources. It is proper to acknowledge the future use of these methods for possible better results.

One such method is the application of a mask over the lung voxel to focus on the tissue within the lung (Figure 4c) where the tuberculosis is located. This mask [21], computes the connected regions of the voxel (Figure 4a), by labeling all connected pixel with the same value. The same mask further computes the lung with no internal structure by computing for every slice the maximum connected pixels, those being the lung frontier if that slice contains lungs, or none otherwise (Figure 4b).



(a) Lung voxel (b) Voxel of lung without internal structures (c) Lung internal structure

Fig. 4: Masks of a lung

3.3 Building the prediction algorithm

Instead of training a binary classifier for severity LOW and HIGH, we instead used the original severity score from 1 to 5 found in the patient's metadata. For the final prediction, we added the probabilities estimated by our model for severity scores 1, 2 and 3 to obtain the probability of HIGH severity.

As previously stated in the introduction, the objective of our paper is to explore methods that minimize the amount of resources used for the prediction. This implies finding a way to reduce the amount of information used during the

prediction. Our idea is to use a model capable of selecting the most important features. We consider each pixel as a feature for our model. In this way, by selecting a subset of the features, we are actually restricting the information we need from the CT to make predictions. Moreover, we would like to supply an upper bound for the number of utilized pixels.

A good choice for such a model is the Decision Tree algorithm described at [18]. A simple variant for limiting the number of nodes is by stopping the algorithm when the number of nodes reaches the threshold. Another method for building decision trees with a maximum number of nodes is provided by [8], where they introduce constraints into the building phase of the algorithm. By setting a size constraint on the decision tree, the best nodes will be selected within that constraint.

Boosting algorithms use a linear combination of weak classifiers to make predictions. They learn the coefficients of each linear classifier in a stage-wise fashion by first training a weak learner and then selecting the weight as to minimize a loss function. If the weak learner utilizes only a limited number of different features, an upper bound of the boosting algorithm is the number of weak learners times the number of different features. This motivates our choice of using a boosting algorithm with decision tree as weak classifiers. To improve on this, consider the small number of instances (<400). By using a boosting algorithm, we try to prevent overfitting while also being able to limit the number of features our algorithm is allowed to use for the prediction.

For the original severity scoring prediction, the ability to supply probabilities is required. A boosting algorithm that can estimate probabilities in a multi-labeling setting is described at [20], called SAMME.R. This particular algorithm is able to provide class probability estimates for the prediction. For the weak learners, we used decision stumps. The major problem that we faced while training the model was overfitting. To mitigate this problem, we used four different approaches.

The first technique we used is limiting the number of features a weak learner trains on. For each boosting iteration, we limited the training of the weak learner to 1,000-10,000 pixels sampled at random from an uniform distribution. By using this method the training runtime was also decreased drastically, making possible to train the classifier in low-resource environments, while also reducing overfitting.

Another method that has been used was regularization of decision stumps. This technique was already applied for the AdaBoost algorithm in [2]. We used the same principle and changed the formula to adapt this method for SAMME.R. Instead of minimizing the weighted training error ϵ_w while searching for the best split, we minimize the following objective function:

$$\epsilon_w + \lambda \cdot (P_w(x < s) \cdot H_w(x < s|Y) + P_w(x >= s) \cdot H_w(x >= s|Y) \qquad (1)$$

where H_w is the weighted conditional entropy, P_w is the weighted probability estimate, λ is the regularization factor and s is the split value.

Our third method consists of splitting the training set further into an initial training set and an extended training set. During training, starting from the initial training set, some samples from the extended training set are added. The sample's weight are then renormalized to accommodate the new samples and the algorithm continues as normal. This method adapted for SAMME.R from [16].

Our final approach was removing weak classifiers. We start with only the decision of the first weak classifier and increasingly add weak learners in the trained order. Then the configuration with the highest score on validation data is chosen.

4 Evaluation

For the ImageCLEFmed Tuberculosis SVR prediction task we submitted a trained model based on the SAMME.R algorithm with 1,600 weak learners, where each weak learner is trained on 10,000 features (pixels and metadata) sampled randomly from an uniform distribution at each iteration.

At the time, we didn't yet conceptualize the methods described in the training section and our algorithm suffered from overfitting. This approach obtained a ROC AUC (receiver operating characteristic area under the curve) score of *0.5692* and accuracy of *0.5556*.

In selecting the hyperparametes and the model we used 5-fold cross-validation and ROC AUC, accuracy as metrics. To select the best model, we looked at each fold and chose the most balanced one in terms of our metrics. These are our local results at the time of our submission:

Table 2: 5-Fold cross-validation of the first run

#	Train ACC	Train AUC	Train ACC B	Test ACC	Test AUC	Test ACC B
0	0.9537	0.9967	0.9653	0.3111	0.6511	0.5333
1	0.9306	0.9913	0.9479	0.4444	0.6284	0.5555
2	0.9827	0.9998	0.9827	0.4090	0.5960	0.6136
3	0.9657	0.9969	0.9657	0.4418	0.7489	0.6511
4	0.9717	0.9979	0.9717	0.4634	0.6238	0.6585

- fold index used for testing
Train ACC - training set accuracy original severity score
Train AUC - training set ROC AUC binary severity
Train ACC B - training set accuracy binary severity
Test ACC - test set accuracy original severity score
Test AUC - test set ROC AUC binary severity
Test ACC B - test set accuracy binary severity

After adding the regularization, validation boosting and pruning of weak learners our scores improved significantly:

Table 3: 5-Fold cross-validation of the second run

#	Train ACC	Train AUC	Train AUC B	Test ACC	Test AUC	Test ACC B
0	0.7572	0.9227	0.8208	0.5555	0.7529	0.6666
1	0.9364	0.9858	0.9421	0.4222	0.5988	0.5777
2	0.9195	0.9902	0.9310	0.5227	0.6714	0.6363
3	0.9142	0.9868	0.9428	0.4883	0.7207	0.6976
4	0.8079	0.9320	0.8474	0.5365	0.6095	0.6097

The regularization factor used was 1 and the number of training samples used at some iteration t was $log\ t/log\ n$, where n is the number of weak classifiers. It can clearly be noticed that the model tested on first fold did not overfit on the training data as badly as the other ones, having the lowest training accuracy and the highest Test ROC AUC score. We also trained the model for 1000 iterations.

5 Conclusion

Recent work on tuberculosis severity prediction has been shown to be progressing, ImageCLEFmed Tuberculosis 2019 edition yielding better results both in terms of ROC AUC, as also in accuracy.

Current findings may be considered a further validation of the fact that boosting, one of the modern machine learning approaches, still suffers from severe overfitting when applied on a very noisy dataset with a significantly higher number of features than training instances. However, by employing various data cleaning and novel regularization techniques, it was demonstrated that there are still improvements that can be made to combat this problem. With each improvement on the subject, the prediction of tuberculosis severity will be also made more accessible in low-resources environments to be used by low-buget medical organizations and small practices.

6 Future work

Besides various pre-processing methods, described in the Architecture section, we can make use of the pixels selected by the model to 3D render a helpful voxel of the lung, with emphasis on the most important regions that contribute to the tuberculosis prediction. The latter visualization of a patient data linked with the low resource-demanding model makes it likely for improvement of the tuberculosis investigations.

References

1. Andries, S., Iftene, A.: Predicting user activities using gps collections. pp. 53–60. Proceedings of the 14th Conference on Human Computer Interaction - RoCHI 2017, Craiova, Romania (09 2017)

2. Bereta, M.: Entropy-based regularization of adaboost. Computer Assisted Methods in Engineering and Science **24**(2), 89–100 (2017)
3. Chelmus, R., Gifu, D., Iftene, A.: Prediction of cryptocurrency market. Proceedings of 19th International Conference on Computational Linguistics and Intelligent Text Processing (Cicling). LNCS, Hanoi, Vietnam (09 2018)
4. 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, Lugano, Switzerland (September 9-12 2019)
5. 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, Avignon, France (September 10-14 2018)
6. Dicente Cid, Y., Mller, H.: Texture-based graph model of the lungs for drug resistance detection, tuberculosis type classification, and severity scoring: Participation in the imageclef 2018 tuberculosis task (09 2018)
7. Gabor, S., Iftene, A.: Mir a system for medical image retrieval. pp. 9–19. Curative Power of Medical Data - Selected Papers of the First International Workshop MEDA 2017, "Alexandru Ioan Cuza" University, Iasi, Constanta, Romania (09 2017)
8. Garofalakis, M., Hyun, D., Rastogi, R., Shim, K.: Building decision trees with constraints. Data Mining and Knowledge Discovery **7**(2), 187–214 (Apr 2003)
9. Iftene, A., Siriteanu, A.: Using semantic resources in image retrieval. 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, vol. 96, pp. 436–445. Elsevier, York, United Kingdom (09 2016)
10. Iftene, A., Vamanu, L., Croitoru, C.: Uaic at imageclef 2009 photo annotation task. pp. 283–286. C. Peters et al. (Eds.): CLEF 2009, LNCS 6242, Part II (Multilingual Information Access Evaluation Vol. II Multimedia Experiments), Springer, Heidelberg (09 2010)
11. 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)
12. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
13. Kovalev, V., Liauchuk, V., Skrahina, A., Astrauko, A., Rosenthal, A., Gabrielian, A.: Examining the utility of clinical, laboratory and radiological data for scoring severity of pulmonary tuberculosis (06 2018)

14. Liauchuk, V., Tarasau, A., Snezhko, E., Kovalev, V.: Imageclef 2018: Lesion-based tb-descriptor for ct image analysis (09 2018)
15. McIntosh, J.: All you need to know about tuberculosis (2019)
16. Meijer, D.W.J., Tax, D.M.J.: Regularizing adaboost with validation sets of increasing size. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 192–197 (Dec 2016)
17. Organization, W.H.: Bcg vaccines: Who position paper february 2018 vaccins bcg: Note de synthse de loms fvrier 2018. Weekly epidemiological record **93**(08), 73–96 (Feb 2018)
18. Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (Mar 1986)
19. Serban, C., Siriteanu, A., Gheorghiu, C., Iftene, A., Alboaie, L., Breaban, M.: Combining image retrieval, metadata processing and naive bayes classification at plant identification 2013. Notebook Paper for the CLEF 2013 LABs Workshop - ImageCLEF - Plant Identification, Valencia, Spain (09 2013)
20. Zhu, J., Rosset, S., Zou, H., Hastie, T.: Multi-class adaboost. Statistics and its interface **2** (02 2006)
21. Zuidhof, G.: Full preprocessing tutorial (2019)