Lung-Wise Tuberculosis Analysis and Automatic CT Report Generation with Hybrid Feature and Ensemble Learning

Muhammad Waqas, Zeshan Khan, Shaheer Anjum, and Muhammad Atif Tahir

National University of Computer and Emerging Sciences, Karachi, Pakistan
{waqas.sheikh,zeshan.khan,k163603,atif.tahir}@nu.edu.pk

Abstract. This article presents the proposed methodology for tuberculosis analysis and generation of the computerized report by using 3D Computed Tomography (CT) images, apropos to the ImageCLEF tuberculosis CT report generation task. The contribution of this paper is based on the combination of handcrafted and non-handcrafted feature extraction strategies. Experiments show that more informative input representation can be obtained by combing different feature extraction strategies that lead to improved performance. In this work, non-handcrafted features mined by using a fine-tuned version of a pre-trained VGG19 model and handcrafted features extracted using Local Binary Pattern (LBP), Haralick, and Intensity Histogram (IH) descriptors. Extracted features combined by using early fusion and final probability estimation performed with an ensemble-based soft voting approach. The proposed methodology achieved a 70.5% mean area under the curve AUC and ranked 6th on the leaderboard for best participation by each group. The proposed approach can be further improved by adopting optimized feature selection and fusion techniques.

1 Introduction

Tuberculosis (TB) is a bacterial disease, it is an airborne disease that attacks the respiratory system, through droplets released by the patients via cough. According to the findings of WHO, tuberculosis caused around 1.3 million deaths in 2017 and 2018 [26]. Timely Diagnosis and treatment of TB can hinder the deaths of patients. The recent advancements in imaging technologies are helping medical practitioners to manually analyze the severity of TB, such as Computed Tomography (CT) scan, which is commonly used for obtaining lesion patterns. In a single CT image, multiple 2D radiographic projections or 2D slices are captured around the objects, and a 3D volume is constructed which allows visualization and slicing at any angle; however, these manual procedures for severity detection are prone to error and costly in terms of time and capital. On the other

hand, machine learning techniques are used for disease analysis which opens new research areas for the researchers. These automatic medical image analysis techniques have shown a proficiency for several imaging modalities, in terms of time and precision [1,4].

In the case of CT images, variation in inter-slice distance, sizes, and shape of voxels entail difficulty in image analysis. Additionally, advanced image analysis algorithms are developed by using deep learning techniques. These algorithms required a large amount of training data and the unavailability of adequate CT imaging data is a major barrier to the use of deep learning systems for automatic tuberculosis analysis.

To alleviate the problem of data unavailability in the domain of medical image analysis, the Cross-Language Evaluation Forum (CLEF) organizes several challenges through the ImageCLEF initiative every year. These challenges aim to provide standard datasets for disease analysis and medical image retrieval [14]. Tuberculosis task was first introduced in 2017's edition [6], and continuously presented every year since then[6,7,8,14], and substantial data for training and testing were provided in these editions. This year the task was to generate an automatic report for detailed lung wise analysis using CT images. The report has to include probability scores for six different class labels[23]. As we know that feature extraction is one of the most challenging part in any machine learning problem. The studies in [20,18] compared the performance of handcrafted and non-handcrafted feature extraction techniques and found that the transfer learning approach for feature extraction performs better than handcrafted feature extraction methods. However, the experiments also demonstrated that both feature extraction strategies obtain dissimilar information from input data, and fusion on these features shown better performance than a single feature extraction strategy. From taking the motivation from [20,18], this paper aims to study both form of feature extraction strategies, handcrafted and non-handcrafted feature extraction, to obtain more informative representation from the input, and their combined impact on classification performance for tuberculosis analysis.

For the CNN based feature extraction, we fine-tuned the VGG19 model [24], pre-trained on the ImageNet dataset [10]. Besides, deep features, we experimented with Haralick [11], LBP [21], and Intensity Histogram (IH) feature extraction techniques for hand-crafted features. Finally, after the fusion of both types of features, the final probability scores for each class is calculated using an ensemble-based soft majority voting approach.

The proposed method provides the benefits of simplicity and generality: The method is less computationally expensive as compared to training deep learning models, which required time and resources. The proposed approach is also useful when the size of available training data is small, and training a deep learning model might not be advantageous. Furthermore, fused descriptors could easily be used to train any classification model for arbitrary labeling.

The organization of the paper is as follows. Section 2 describes the dataset, the types of images, and the splitting criteria for train and test distribution.

Section 3, discusses the proposed methodology. Section 4 presents the results of the experiments. Finally, We make conclusion and present potential future work.

2 Task and Dataset Description

The tuberculosis task in previous editions of ImageCLEF was divided into several subtasks, such as severity scoring, TB types detection and CT report generation. However, The objective of this year's task is to generate an automatic lung-wise report that incorporates probability scores for six different class labels, including "Left-Lung-Effected", "Right-Lung-Effected", "Caverns-Left", "Caverns-Right", "Pleurisy-Left" and "Pleurisy-Right" respectively, based on the CT image data [13,12,17].

The dataset consists of 3D CT images in NIfTI (Neuroimaging Informatics Technology Initiative) format and stored with the ".nii.gz" extension. Each 3D CT image compromised of around 100 2D slices of size 512*512. In this year's edition, the dataset consists of 403 3D CT images, further divided into 283 training and 120 testing images. The dataset is labeled lung-wise, which double the size of training examples for lung-wise analysis. The numbers of occurrence for each class label in training data are shown in Table 1.

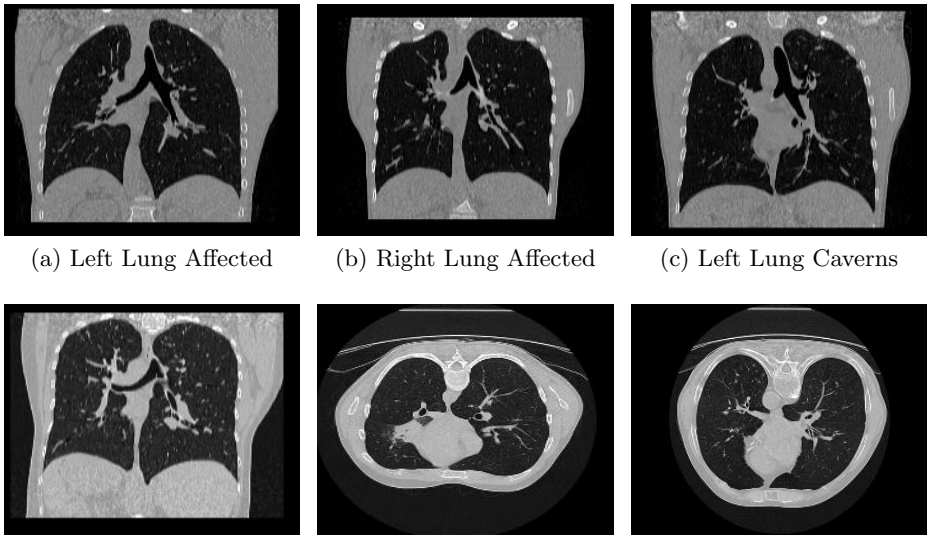
Furthermore, an automatically extracted lung mask is also provided for every patient [5,19]. The numbers of occurrence for each class label in training data are shown in Table 1. Furthermore, some of the image slices are shown in figure 1.

Table 1: Class Distribution in Tuberculosis Dataset [12]

Sr#	**Label**	**Number of Occurrences**
1	Left-Lung-Effected	211
2	Right-Lung-Effected	233
3	Caverns-Left	66
4	Caverns-Right	79
5	Pleurisy-Left	7
6	Pleurisy-Right	14

3 Methodology

In this section, the proposed methodology is discussed in detail. The methodology is a 4 stage process, which includes preprocessing, feature extraction, fusion, and finally classification. All these stages are discussed in detail and shown in figure 2.



(a) Left Lung Affected (b) Right Lung Affected (c) Left Lung Caverns

(d) Right Lung Caverns (e) Left Lung Pleurisy (f) Right Lung Pleurisy

Fig. 1: Images of Six Different Classes from ImageCLEF 2020 Tuberculosis Dataset

3.1 Preprocessing

The proposed approach intends to fuse several features from each slice, for this purpose the provided NIfTI format images are first converted into .png format by using NiBabel library [9]. The conversion is accomplished by extracting all slices of size 512*512 and stored in .png format [3], and around 100 images in .png format are extracted from each 3D CT image.

3.2 Fine-tuning Pre-trained VGG19

Deep learning models require considerably large training time and training data to achieve good results, however, this necessity can be alleviated using transfer learning. In this approach, a complex representation previously learned from a large training dataset by a model, which can be reused as input for a second task. This approach has shown remarkable performance in several medical image Classification frameworks [16,15,22]. For the process of transfer learning pre-trained VGG19 model [24] trained on ImageNet data[10] is fine-tuned. The pre-trained model modified by substituting the last three layers which are defined for the ImageNet dataset, by three fully connected layers of 1024,512 and 6 neurons respectively. The modified network is then retrained by using Stochastic Gradient Descent (SGD), by fixing a learning rate, momentum, and a mini-batch size to 0.01, 0.9 and 30 respectively; moreover, 50 epochs are performed for each

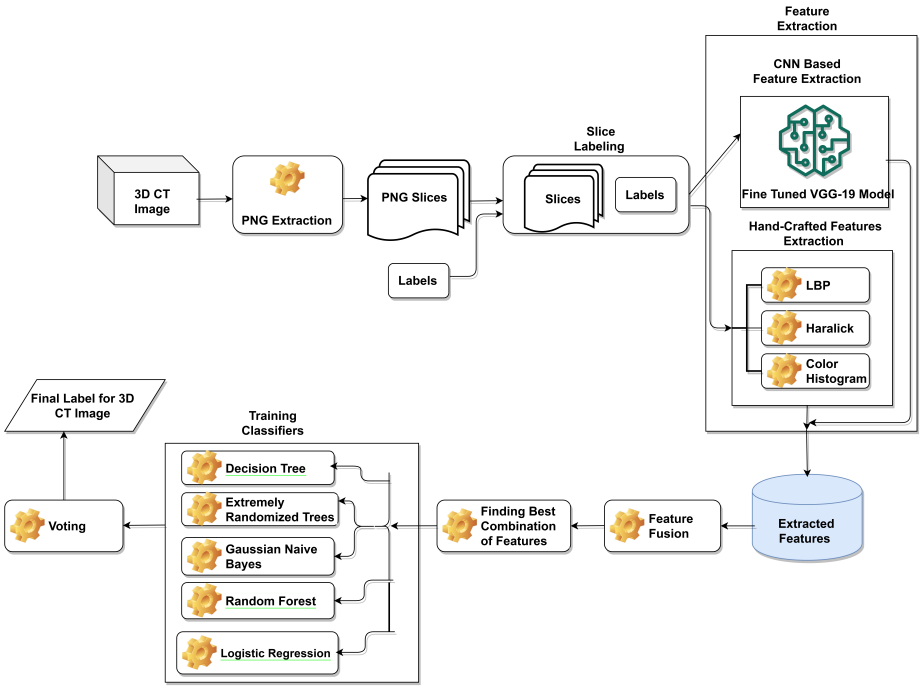


Fig. 2: Proposed System Architecture

provided part of the dataset i.e., the original, masks1, and masks2. The dropout rate of neurons and weight decay parameters are used to avoid overfitting in a predefined network.

3.3 Feature Extraction and Fusion

The extracted features are combined using early fusion technique, combinations of various features are evaluated in comparison with deep features. To validate the performance of each combination of descriptors, we used various classifiers, including Decision Tree [25] (DT), Extremely Randomized Tree [2] (ET), Random Forest (RF), Logistic Regression (LR) and Gaussian Naive Bayes (GNB), for evaluation criteria we used average F1-score.

The training data is divided into two parts, training part and validation part with a ratio of 75% and 25% respectively. Based on experiments we selected (LBP and Haralick) features beside deep features for further experiments. The performance of several combinations of features is presented in Table 2.

Table 2: Performance of Different Descriptor

Sr#	**Features Combination**	**Average F-Score**
1	LBP	0.70
2	Haralick	0.73
3	IH	0.72
4	Deep Features	0.91
5	LBP + Haralick + Deep Featues	0.95

3.4 Classification

For classification, ensemble based strategy is adopted. We trained DT,LR and GNB classifiers independently trained on hybrid feature vectors, and final results were combined with soft and hard voting techniques.

The ensembles of classifiers can have hard and soft voting. Hard voting counts the vote or predicts Y the label through majority predicted class based on equation 1, here C_m is the predicted class label of model m. Soft voting predict the class label by using predicted probability P_c by each classifier based on equation 2 where W_c is assigned weight to c^{th} classifier.

$$Y = \text{mod}\{C_i(x) : i \in models\} \tag{1}$$

$$Y = argmax \sum_{j=1}^{c} W_j P_{i,j} \tag{2}$$

The resulted probability scored for each image-slice are then passed to a threshold function to obtain final class label, described in equation 3 where P_i is the probability of i^{th} class label.

$$P_i = \begin{cases} 1, & \text{if } x \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Finally, the ultimate probability scores for of single 3D CT image is computed by the averaging the class labels for all the image-slices inside a single CT image as shown in equation 4, where P_j is the probability score of j^{th} class and S_k is number of slices in each 3D image. We used scikit-learn package implementation for classification models. Hyperparameters for all of the models were tuned by cross-validation using grid search.

$$P_j = (1/S_k) \sum_{i=1}^{k} C_{i,j} \tag{4}$$

4 Submission and Results

The method described in the previous section was applied to generate predictions for the test set[12,17]. The labels for test set were not provided, all the

participations were evaluated using AUC, and final results and the participant standing were calculated by organizers. We submitted three different runs, details of each run is given below. A complete list of the results for the task is available at ImageCLEF Website.

- **Run1** In this run, the results are obtained by training ensemble model discussed in section 3.4 using best performing combination of descriptors (LBP + Haralik + Deep features), and with the usage of soft voting approach.
- **Run2** In this run, the ensemble model is trained on the descriptors as in Run 1 and the hard-voting technique is applied instead of soft voting.
- **Run3** In this run, the ensemble model is trained on the fused version of descriptors obtained by using all features extraction techniques, mentioned in section 3.3, and performance is tested by applying soft-voting technique.

Tables 3 show the details of the best results achieved by the participating group, Our group **FAST_NU_DS** ranked 6th. Our best-submitted run achieved mean AUC of 0.705, and minimum AUC of 0.644. Table 4 shows the performance of each submitted run in detail.

Table 3: Top 9 Results of ImageCLEF 2020

Group Rank#	**Group ID**	**Mean AUC**	**Minimum AUC**
1	SenticLab.UAIC	0.924	0.885
2	SDVA-UCSD	0.875	0.811
3	chejiao	0.791	0.682
4	CompElecEngCU	0.767	0.733
5	KDE-LAB	0.753	0.698
6	**FAST_NU_DS**	0.705	0.644
7	uaic2020	0.659	0.562
8	JBTTM	0.601	0.432
9	sztaki_dsd	0.595	0.546

Table 4: All Three Submissions

Run #	**Submission ID**	**Mean AUC**	**Minimum AUC**	**Submission Rank**
Run 1	67947	0.705	0.644	37
Run 2	68125	0.567	0.458	52
Run 3	68128	0.496	0.481	58

The best results are obtained by Run 1 followed by Run 2 and Run 3. It can be observed that hard and soft voting techniques can lead to dissimilar decision boundaries.

Run1 with soft voting shows the best performance, since it takes into classifier's uncertainties in the final decision, and the final decision boundary relies on strong classifier and works well when classifiers are carefully adjusted. Furthermore, incorporating only important features in classification removes redundancy in input space and helps to reduce the complexity of learner, Due to this, a clear difference can be seen in the performance of Run 1, Run 3. As compared to Run 1, the performance of Run 3 suffers from the redundancy in input space. The results obtained by our submitted runs are not well ranked as compared to the top-ranked runs. This is due to the fact that each team has submitted several runs and performance variation between them is probably not high.

5 Conclusion and Future Work

In this article, we presented our contribution to ImageCLEFmed 2020 Tuberculosis task. We used the combination of transfer learning and handcrafted feature extraction techniques. In the proposed approach, VGG19 model fine-tuned for transfer learning and extracted features are fused with LBP and Haraclick features. Results show that two different feature extraction methods can obtain diverse representation for input, and performs better as compared to the standalone feature extraction approach. Moreover, an ensemble-based soft voting approach is proposed for the classification of 3D CT images. The proposed technique is simple, less resource-oriented, but yet effective. Although the proposed technique has not produced the best result, however, the performance of the proposed technique could be further improved by combing several other deep and handcrafted features and adopting some optimized way to select the set of best performing attributes from the fused vector. Furthermore, In future work, heuristic strategies for sample selection and feature selection will be adopted. Additionally, sieving technique to select informative slices from a 3D image, and ignoring unnecessary slices or slices with no information will also be explored. This could lead to further improvement in performance.

References

1. Yousef Al-Kofahi, Alla Zaltsman, Robert Graves, Will Marshall, and Mirabela Rusu. A deep learning-based algorithm for 2-d cell segmentation in microscopy images. *BMC bioinformatics*, 19(1):1–11, 2018.
2. Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.
3. P Atzberger. Portable network graphics. *Web Tech.*, 1:65–68, 1996.
4. Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.
5. Yashin Dicente Cid, Oscar Alfonso Jiménez Del Toro, Adrien Depeursinge, and Henning Müller. Efficient and fully automatic segmentation of the lungs in ct volumes. In *VISCERAL Challenge@ ISBI*, pages 31–35, 2015.

6. Yashin Dicente Cid, Alexander Kalinovsky, Vitali Liauchuk, Vassili Kovalev, and Henning Müller. Overview of the imageclef 2017 tuberculosis task-predicting tuberculosis type and drug resistances. In *CLEF (Working Notes)*, 2017.
7. Yashin Dicente Cid, Vitali Liauchuk, Dzmitri Klimuk, Aleh Tarasau, Vassili Kovalev, and Henning Müller. Overview of imagecleftuberculosis 2019-automatic ct-based report generation and tuberculosis severity assessment. In *CLEF (Working Notes)*, 2019.
8. Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, and Henning Müller. Overview of imagecleftuberculosis 2018-detecting multi-drug resistance, classifying tuberculosis types and assessing severity scores. In *CLEF (Working Notes)*, 2018.
9. Thiago F De Moraes, PH Amorim, Fábio S Azevedo, and JV da Silva. Invesalius—an open-source imaging application. *Comput Vis Med Image Process*, page 405, 2011.
10. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
11. Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.
12. Bogdan Ionescu, Henning Müller, Renaud Péteri, Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, Serge Kozlovski, Vitali Liauchuk, Yashin Dicente Cid, Vassili Kovalev, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca PImageCLEF20iras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Duc-Tien Dang-Nguyen, Jon Chamberlain, Adrian Clark andImageCLEF20 Antonio Campello, Dimitri Fichou, Raul Berari, Paul Brie, Mihai Dogariu, Liviu Daniel Ştefan, and Mihai Gabriel Constantin. Overview of the ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260 of *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*, Thessaloniki, Greece, September 22-25 2020. LNCS Lecture Notes in Computer Science, Springer.
13. Bogdan Ionescu, Henning Müller, Renaud Péteri, Duc-Tien Dang-Nguyen, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, et al. Imageclef 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In *European Conference on Information Retrieval*, pages 533–541. Springer, 2020.
14. Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at imageclef 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015.
15. Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
16. Zeshan Khan and Muhammad Atif Tahir. Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract. In *MediaEval*, 2018.

17. Serge Kozlovski, Vitali Liauchuk, Yashin Dicente Cid, Aleh Tarasau, Vassili Kovalev, and Henning Müller. Overview of ImageCLEFtuberculosis 2020 - automatic CT-based report generation. In *CLEF2020 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece, September 22-25 2020. CEUR-WS.org <http://ceur-ws.org>.
18. Sheng Long Lee, Mohammad Reza Zare, and Henning Muller. Late fusion of deep learning and handcrafted visual features for biomedical image modality classification. *IET image processing*, 13(2):382–391, 2018.
19. Vitali Liauchuk and Vassili Kovalev. Imageclef 2017: Supervoxels and co-occurrence for tuberculosis ct image classification. In *CLEF (Working Notes)*, 2017.
20. Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172, 2017.
21. Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:971–987, 2002.
22. Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.
23. Yashin Dicente Cid Aleh Tarasau Vassili Kovalev Serge Kozlovski, Vitali Liauchuk and Henning Müller. Overview of imagecleftuberculosis 2020 - automatic ct-based report generation and tuberculosis severity assessment. In *CLEF (Working Notes)*, 2020.
24. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
25. Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.
26. WHO. *World Health Organization*, 2020 (accessed April 7, 2020).