Ensemble of Deep Learning Models for Automatic Tuberculosis Diagnosis Using Chest CT Scans: Contribution to the ImageCLEF-2020 Challenges

Abdela A. Mossa[1][0000−0002−6168−5002], Halit Eriş[2][0000−0002−2384−5052], and Ulus Çevik[2][0000−0002−0956−9725]

[1] Department of Computer Engineering, Çukurova University, Adana, Turkey
[2] Department of Electrical-Electronics Engineering, Çukurova University, Adana, Turkey (amossa@student., heris@, ucevik@)cu.edu.tr

Abstract. Tuberculosis (TB) is a bacterial infection that mainly affects the lungs. It is a potentially serious disease killing around 2 million people a year. Nevertheless, it can be cured if treated with the right antibiotics. However, manual diagnosing of TB can be difficult, and several tests are usually conducted by clinicians. Consequently, automated diagnosis of TB based on chest Computed Tomography (CT) images for rapid and accurate diagnosis are currently of great interest. Recently, deep learning algorithms, and in particular convolutional neural network (CNN), due to the ability to learn low- and high-level discriminative features directly from images in an end-to-end architecture, have been shown to be the state-of-the-art in automatic medical image analysis. In this work, we developed a deep learning model for automated TB diagnosis using an ensemble of different CNN architectures trained on 2D images sliced from volumetric chest CT scans. The CNN-based methods proposed in this study includes Multi-View and Triplanar CNN architectures using pre-trained AlexNet, VGG11, VGG19 and GoogLeNet feature extraction layers as a backend. Using five-fold cross validation, the average AUC, Accuracy, Sensitivity and Specificity of the proposed ensemble method were 0.799, 77.1, 0.57 and 0.824, respectively, for multi-label binary classification on the ImageCLEFtuberculosis 2020 training dataset of the lung-based automated CT report generation task, which is a well-benchmarked public dataset running every year since 2017. The result shows the strength of our model trained in a small dataset with highly unbalanced label distributions, leading to 4^{th} place on the Leaderboard, with a mean AUC of 0.767 on the test dataset.

Keywords: Automatic CT Report Generation · Deep Learning · Convolutional Neural Network · Tuberculosis Diagnosis · 3D Medical Image Analysis.

1 Introduction

Tuberculosis (TB) is a highly contagious disease that typically attacks the lungs. Every year, approximately ten million people become infected with TB, with around one and half million deaths, thereby making the disease a global health problem [1]. Even though many researches have been done to reduce the spread of TB in the society, the report by the World Health Organization (WHO) in 2019 [2] indicates that TB still remains at the top ten causes of death worldwide and epidemic in 202 countries and territories (see Table 1).

Table 1. Number of countries and territories that reported the TB incidents to WHO in 2019.

Regions	Numbers
Africa	46
European	45
Region of the Americas	43
Western Pacific	35
Eastern Mediterranean	22
South-East Asia	11
Global	202

Computed Tomography (CT) is one of the most commonly used non-invasive medical imaging techniques in the diagnosis and management of patients with TB [3]. A volumetric chest CT scan of people with suspected TB is obtained and examined either for abnormalities suggestive of TB or for detection of any kind of TB abnormality. It aids physicians to visualize lesions with specific manifestations in the altering lung tissues caused by tuberculosis [4]. However, CT comes at the cost of generating thousands of images per patient, which makes it time-consuming, subjective, and even impossible to achieve high performance level in the absence of expert radiologists [5]. Hence, the development of computer-aided diagnosis (CAD) techniques to assist physicians in tuberculosis detection and diagnosis have been attracted much attention from researchers at the intersection of medicine and artificial intelligence [6–9].

Deep learning (DL) [10] based CAD algorithms especially convolutional neural networks (CNNs) [11] that learn visual patterns directly from images with minimal pre-processing and without the intermediate step of experts have recently been effective in the medical imaging and other computer vision applications [12–14]. Along these lines, as part of CLEF (Conference and Labs of the Evaluation Forum) - a series of campaigns that have been carried out in the information retrieval domain since 2000, ImageCLEF 2020 has presented an evaluation campaign that offers researchers around the world to participate in the ImageCLEFtuberculosis task that runs for fourth consecutive year [15, 16].

The task provided by ImageCLEFtuberculosis organizers varies from year to year. Last year the tasks were Severity Score Prediction (SVR) and CT – based

automatic CT report generation (CTR) based on volumetric chest CT scans and clinical information of patients [17]. However, this year's challenge (ImageCLEF-tuberculosis 2020) was a lung-based automatic CT report generation solely on CT images [18]. In last year's tuberculosis challenge, even though we participated for the first time, our Multi-View CNN based approach achieved rank 4^{th} with mean AUC of 0.707 [19]. Hence, since our last year approach produced competitive result, we decided to improve and adapt it to the requirements of this year challenge. Therefore, in this study, we developed a novel CAD based system for automated TB diagnosis by using different Multi-View and Triplanar CNN architectures with the ensemble method on chest CT images. We developed the CNN architectures using pre-trained AlexNet [20], GoogLeNet [21], and VGG [22] feature extraction layers as a backend.

This paper has the following structure: in section 2, we present the dataset, image pre-processing, CNN architectures and ensemble methods used in this work. Results and discussions are reported in section 3. Finally, section 4 points future works, and concludes this paper.

2 Materials and Methods

2.1 Dataset and Image Pre-processing

The training and test datasets provided by the ImageCLEFtuberculosis 2020 task organizers consist of 403 studies of people with TB where the organizers divided the dataset into 283 training and 120 test studies. Each study contains patients' volumetric chest CT scans stored in NIFTI file format, automatically extracted masks of the lungs obtained with the algorithm discussed in [23], and a lung-based six diagnosis labels, which are:

(i) LeftLungAffected (LL) - binary label for presence of any TB lesions in the left lung;

(ii) RightLungAffected (RL) - binary label for presence of any TB lesions in the right lung;

(iii) CavernsLeft (CL) - binary label for presence of caverns in the left lung;

(iv) CavernsRight (CR) - binary label for presence of caverns in the right lung;

(v) PleurisyLeft (PL) - binary label for presence of pleurisy in the left lung;

(vi) PleurisyRight (PR) - binary label for presence of pleurisy in the right lung.

The provided training dataset by the task organizers is highly imbalanced in which there are more positive cases than negative cases in LL and RL labels, and few positive cases than negative cases in the other diagnosis labels. Moreover, the PL label has the largest unbalanced distribution in the dataset where the proportion of positive training cases being about 2.5%. Even the CVR label, which has a relatively better balanced distribution than the other labels, has only 27.9% of the training cases labelled positive. Fig.1 depicts the number of positive and negative patients for each of the six diagnosis labels of the training dataset. More details about the datasets can also be found in [17].

The sizes of all the volumetric chest CT scans are $512 \times 512 \times k$, where image length and width are 512 and k indicates number of slices in the axial plane varying from 47 to 264 and 101 to 258 for training and testing datasets, respectively. We used the training dataset to develop a model that can generate multi-class binary classification prediction results related to the three labeled diagnosis conditions of each lung. In other words, our model simultaneously predicts whether a certain condition is present (i.e. 'positive or the numerical equivalent of 1') or absent (i.e. 'negative or the numerical equivalent of 0') for each of the three diagnosis labels of each lung.

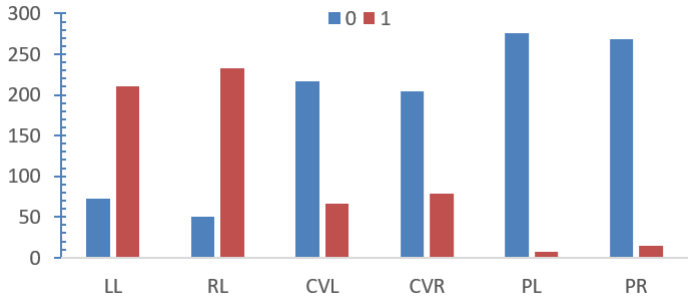

Fig. 1. Distribution of the positive and negative cases across the different diagnosis labels. For LL and RL, the majority of cases are "Positive" compared to the minority of "Negative" cases. However, for CVL, CVR, PL and PR labels the majority of cases are "Negative".

As we planned to leverage 2D CNN models pre-trained on natural images of a fixed image resolution, we reformatted each 3D chest CT scan to a group of 2D stacked slices in the axial, coronal and sagittal views, respectively. Each axial slice is then cropped to a fixed size of 256×256 pixels around the left and right lung regions, respectively. Similarly, we cropped each sagittal and coronal slices to a fixed size of 128×256 pixels around the left and right lung regions, respectively. The rectangular bounding box locations around each lung were selected through visual inspection of few mid-level slices using the provided segmented masks. To avoid processing the background which does not contain any lung tissue and process the scans under the memory constraints of the GPU, only 30 axials, 60 coronal and 60 sagittal mid-level slices from each volumetric chest CT exams were selected. In addition, to avoid the effect of image enlarging on the models classification performance, two consecutive sagittal slices and two consecutive coronal slices, respectively, were concatenated and reshaped to 256×256 pixel sizes. Then, we rescaled the intensity values of the slices to (0,255) range, convert them to PNG format, and normalized to have zero mean and unit variance. Then, all the sliced axial, sagittal and coronal PNG images were

stacked together, and saved in serialized form with pickle toolbox, respectively. Therefore, our input shape turned to be (30, 3, 256, 256). The values can be interpreted such that first value holds for the number of axial, coronal or sagittal slices after pre-processing. The last two values for width and height of images and 3 represents the number of color channels. The sketch map of image pre-processing steps is shown in Fig.2

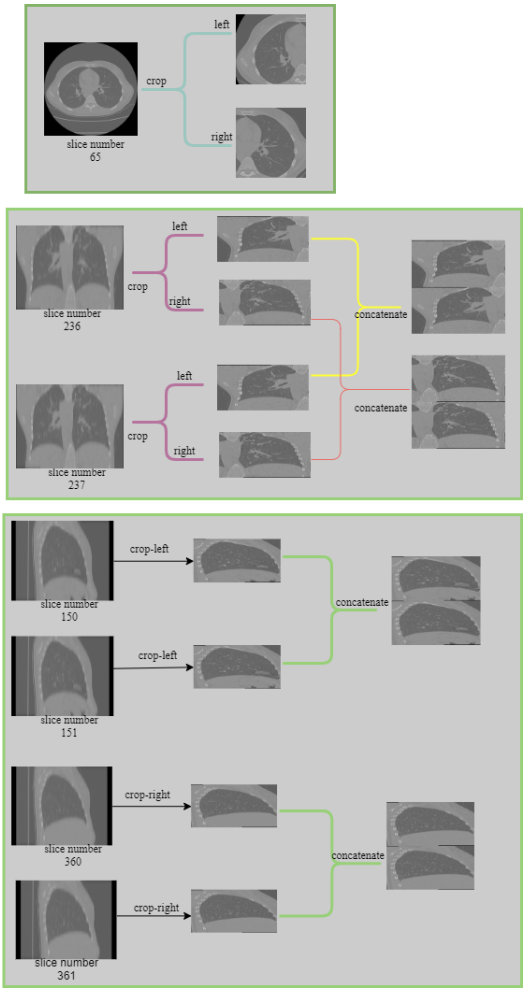

Fig. 2. Example of 3D chest CT scan of patient ID CTR-TRN-051 pre-processing stages. From top to bottom: 2D sliced from the 3D scan and then cropped around the left and right lung regions in the axial, coronal, and sagittal views, respectively.

2.2 Model Development

Convolutional neural network (CNNs), also known as deep learners are machine learning methods designed to process image data via convolutional, pooling and fully connected layers. Convolution and pooling layers occur in an alternative fashion to extract high-level features, and fully connected layers are used to perform classification. In this paper, we aimed to develop a DL model that simultaneously predicts lung-based TB diagnosis labels by using different CNN architectures with the ensemble method on chest CT images. We address it as a multi-class binary classification problem. Moreover, we repeated training the proposed architectures two times, one for each lung related diagnosis labels report generation task.

Considering the training dataset being very small and heavily imbalanced, we proposed five CNN architectures (3 Multi-View CNN architectures: AlexNet$_{MV}$, GoogLeNet$_{MV}$ and VGG19$_{MV}$, and 2 Triplanar-CNN architectures: AlexNet$_{TP}$ and VGG11$_{TP}$) using pre-trained AlexNet, GoogLeNet, VGG11, and VGG19 feature extraction layers as a backend. All of the five CNN architectures were trained using Adam optimization with backpropagation algorithms as they are successfully applied in many deep learning models. In addition, all the models were optimized using weighted binary cross-entropy loss function to account for the unbalanced class sizes. The parameters tuning were experimentally determined individually for each proposed architecture. Moreover, when the validation loss did not decrease for 20 epochs, we early-stopped the parameter optimization and training process to avoid the overfitting problem. Then, the model with the lowest average loss on the validation dataset were selected as our final model candidate. All the models were developed by using a desktop computer with NVIDIA GeForce RTX 2070 GPU and the widely used deep learning framework Pytorch with backend libraries of Tensorflow [24].

The individual classification performance of the five CNNs on the training and testing datasets were compared. Then, in order to get a better and more comprehensive generalized model [25], and motivated by the idea of "two or more heads are better than one", the probability predictions by the four CNNs that performed better were fused using different strategies: average, majority voting and stacking (Naïve Bayes). The probability predictions by GoogLeNet$_{MV}$ was relatively not good compared to the other architectures. Hence, we used the other four CNN models as our base learners in the ensemble approach we used. Details of each model architecture and results are discussed in the following parts.

Model Architectures

Multi-View CNNs. The Multi-View CNN architectures proposed in this work are an extension of our prior work for last year year's TB challenge [19]. In the paper, coronal and sagittal slices were concatenated before fed to the AlexNet based multi-view CNN model, and axial slices were not used. However, in this work's proposed Multi-View CNN , in addition to AlexNet, we used pre-trained

VGG19 and GoogLeNet feature extraction layers as a backend. Moreover, in addition to coronal and sagittal slices, axial slices is also used in this work to train the proposed models.

The basic concept of the proposed Multi-View CNN architecture is that during the training process we provide the model a serious of 2D axial images sliced from 3D CT scan as input and similar sagittal and coronal images as data augmentation techniques, and generates a classification prediction results for each lung related labels. As depicted in Fig.3, the overall Multi-View CNN architecture consists of three core parts:

(i) The feature extraction layers of pre-trained state-of-the-art CNN model (i.e VGG19, AlexNet or GoogLeNet).

(ii) Global average pooling and max pooling layers on top of the feature extraction layers applied across the spatial dimensions to reduce feature maps, and

(iii) Dense layer. The dense layer was fed the resulted feature maps after pooling operations. Then, the sigmoid function applied to the output of the dense layer to obtain the final probability binary prediction score for each of the three diagnosis labels of each lungs.

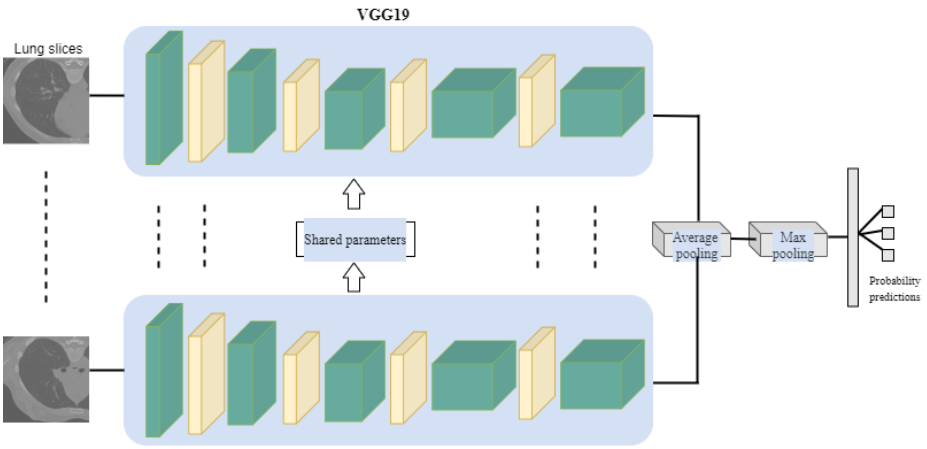

Fig. 3. Multi-View CNN architecture: VGG19$_{MV}$. VGG19$_{MV}$ is an automatic TB diagnosis Multi-View CNN architecture using VGG19 feature extraction layers as a backend. The architecture takes as input a series of CT slices and outputs a multi-class binary classification predictions of the CT scan. Global average and max-pooling operation were used to combine features from each slice obtained using the VGG19 feature extraction layers. The resulted feature maps were then fed to a fully connected layer to generate a probability score of each the three diagnosis labels. We trained VGG19$_{MV}$ two times, one for each lung related report generations. Using similar architecture and training, we developed AlexNet$_{MV}$ and GoogLeNet$_{MV}$ with AlexNet and GoogLeNet feature extraction layers, respectively.

Triplanar-CNN. The overview of the proposed Triplanar-CNN architecture is depicted in Fig.4. A 2D images sliced from the volumetric chest CT scans in the axial, coronal and sagittal planes were fed into the three parallel channels of the Multi-View CNN architecture, respectively. Generated features from the three channels were consolidated into a fixed size feature map to form a single combined feature representation. Then, the classification is performed using a fully connected layer and a sigmoid activation function on top of it. More details on the Triplanar-CNN architecture is available in our prior work developed for automated brain tumor grading [26].

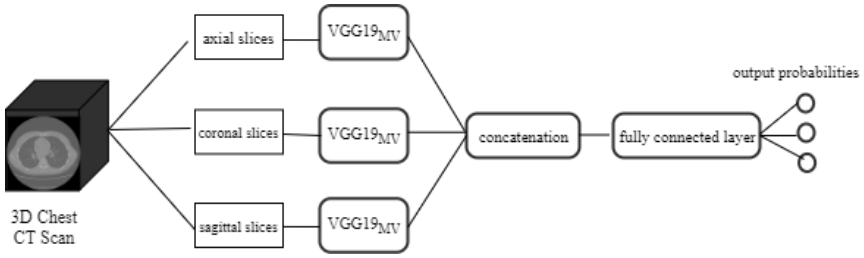

Fig. 4. Triplanar-CNN: VGG19$_{TP}$. A 3D chest CT scan is first decomposed into 2D axial, coronal and sagittal cross-sectional slices then passed to each of the three column VGG19$_{MV}$ Multi-View CNN feature extraction layers, respectively. We trained the VGG19$_{TP}$ two times, one for each lung related report generation. Using similar architecture and training procedure, we developed AlexNet$_{TP}$ using AlexNet$_{MV}$ feature extraction layers in each of the three columns.

3 Results and Discussion

As previously mentioned in Section 2.1, the ImageCLEFtuberculosis 2020 dataset was provided with training and test set partitions. The training dataset is highly imbalanced in each diagnostic labels. Thus, we used five-fold stratified cross-validation upon the training dataset to reduce overfitting and avoid bias during the overall system evaluation in the test dataset. That is, for each validation fold in the training dataset, the remaining other folds were used to train the models. Indeed, this procedure ensures that every CT scan in the training dataset gets to be in the validation set exactly once. The independent testing dataset was not used during training and internal validation. In fact, diagnosis labels of the patients in the test dataset were not visible to the challenge participants. Participants of the challenge were required to submit the probability prediction for each diagnostic labels and ranking was based on the average and minimum AUC over the six diagnostic labels of both the left and right lungs. However, to quantitatively evaluate the capability of the proposed deep learning based approach on both the provided training and testing datasets, the performance

measures averaged over all the five folds of the training dataset are reported in this paper, including the area under the receiver operating characteristic curve (AUC), precision (PRE), specificity (SPE), and sensitivity (SEN) evaluation metrics. Accuracy is not significant for evaluating the performance of the proposed approach as the dataset for each diagnostic labels are highly unbalanced. Performance of our proposed system on both the training and test dataset is explained in the following subsections.

3.1 Performance of the Five Multi-View and Triplanar-CNN Models

Table 2 and 3 reports multi-class binary classification performance of the proposed Multi-View CNN and Triplanar-CNN models, respectively, for both the left and right lung related diagnosis labels, on the training dataset using the five-fold cross validation. The results show that the AlexNet$_{MV}$ classifier achieved better classification performance compared to the other classifiers in terms of mean AUC. Moreover, the AlexNet$_{MV}$ methods outperformed its corresponding Triplanar-CNN model, i.e. AlexNet$_{TP}$, with a marginal increment of 1.6% in terms of mean AUC. Meanwhile, the AlexNet$_{MV}$ classifier outperformed the VGG19$_{MV}$ and the VGG11$_{TP}$ models with improvement rates of 3.2% and 2.6% in terms of the mean AUC, respectively.

GoogLeNet$_{MV}$ suffers with the overfitting problem and it performance (mean AUC of 0.506) was relatively poor compared to the other models. This may be due to the architecture is deeper than the AlexNet and VGG architectures, and due to the scarcity of the available training dataset. In addition to axial slices, Multi-View CNN classifiers were trained using coronal and sagittal slices as data augmentation techniques. However, validation and testing were performed using axial slices only. Triplanar-CNN models were trained and evaluated using axial, coronal and sagittal slices without using any data augmentation techniques. Yet due to the strong performance of the the proposed models, as reported in Table 2 and 3, for the multi-class binary classification problems across the multiple tasks, we are confident that our models will perform better if we were to incorporate extensive data augmentation techniques. In addition, though we used weighted cross-entropy loss to account for the imbalanced class sizes, the performances of the proposed models on some tasks are highly biased towards the majority class. For instance, as shown in Fig.1, out of 283 patients of the training dataset, only 7 (2.5%) of them were PL positive, whereas the remaining 276 (97.5%) were PL negative. Hence, performance of all the models in terms of SEN for the PL binary classification is very poor, whereas the PRE is obviously very high. Similarly, only 4.9% of the training dataset were PR positive, the remaining 95.1% were PR negative. However, unlike that of the PL task, classification performance of all the models in terms of SEN for PR was not highly affected. This shows that the weighted loss computation we used during training the models for tackling imbalanced class size problems worked well for some tasks. Hence weighted loss computation along with some renowned resampling techniques might be further

investigated in order to balance of the classes distribution and avoid bias on classification performance of deep learning models.

Table 2. Performance evaluation of the three Multi-View CNN models. Bold indicates our best results averaged across the six labels.

Models		Left Lung			Right Lung			Avg.
		LL	CVL	PL	RL	CVR	PR	
AlexNetMV	AUC	0.744	0.775	0.82	0.736	0.717	0.88	**0.779**
	SEN	0.609	0.663	0.278	0.644	0.594	0.806	**0.599**
	SPE	0.836	0.799	0.932	0.779	0.746	0.925	**0.836**
	PRE	0.943	0.561	0.194	0.936	0.482	0.387	**0.584**
VGG19MV	AUC	0.76	0.751	0.74	0.71	0.65	0.869	0.747
	SEN	0.632	0.535	0.17	0.656	0.57	0.611	0.529
	SPE	0.71	0.728	0.932	0.714	0.598	0.93	0.769
	PRE	0.923	0.587	0.111	0.92	0.369	0.375	0.548
GoogLeNetMV	AUC	0.566	0.526	0.456	0.454	0.548	0.486	0.506
	SEN	0.591	0.594	0.433	0.79	0.742	0.306	0.576
	SPE	0	0.371	0.383	0	0.46	0.563	0.296
	PRE	0.782	0.265	0.031	0.824	0.35	0.063	0.386

Table 3. Performance evaluation of the two Triplanar-CNN models. Bold indicates our best results averaged across the six labels.

Models		Left Lung			Right Lung			Avg.
		LL	CVL	PL	RL	CVR	PR	
AlexNetTP	AUC	0.706	0.789	0.745	0.697	0.698	0.944	**0.763**
	SEN	0.668	0.646	0	0.665	0.619	0.806	**0.567**
	SPE	0.553	0.892	0.833	0.698	0.737	0.931	**0.774**
	PRE	0.85	0.677	0.083	0.914	0.475	0.446	0.574
VGG11TP	AUC	0.745	0.79	0.69	0.681	0.656	0.957	0.753
	SEN	0.617	0.729	0.0389	0.615	0.553	0.72	0.546
	SPE	0.684	0.741	0.483	0.695	0.714	0.956	0.712
	PRE	0.891	0.582	0.059	0.906	0.45	0.595	**0.581**

3.2 Performance of Ensemble Multi-View and Triplanar-CNN Models

With regard to the AUC, SEN, SPE and PRE, the classification results achieved by each of the three ensemble methods used in our work are reported in Table 4. The mean AUC, SEN, SPE, and PRE of the average fusion strategy were 0.799, 0.571, 0.824, and 0.576, respectively. The mean AUC, SEN, SPE, and PRE of the voting fusion strategy were 0.777, 0.574, 0.821, and 0.574, respectively. The

mean AUC, SEN, SPE, and PRE of the Naïve Bayes fusion strategy were 0.759, 0.573, 0.801, and 0.829, respectively. Average fusion strategy has the highest mean AUC and SPE values, and Naïve Bayes has the lowest in both evaluation metrics. However, Naïve Bayes has the highest mean PRE values than average and voting fusion approaches with improvement rates of more than 25%. The SEN and SPE of the three ensemble methods are almost the same with less than 0.5% and 2.5% difference, respectively.

Table 4. Performance of the proposed system for three different fusion strategies. Bold indicates our best results averaged across the six labels.

Models		Left Lung			Right Lung			Avg.
		LL	CVL	PL	RL	CVR	PR	
Average	AUC	0.788	0.815	0.841	0.73	0.689	0.93	**0.799**
	SEN	0.639	0.614	0.286	0.643	0.574	0.667	0.571
	SPE	0.778	0.832	0.932	0.724	0.721	0.956	**0.824**
	PRE	0.914	0.562	0.154	0.918	0.443	0.462	0.576
Voting	AUC	0.779	0.8	0.736	0.73	0.682	0.936	0.777
	SEN	0.632	0.614	0.286	0.65	0.596	0.667	**0.574**
	SPE	0.75	0.84	0.938	0.724	0.721	0.95	0.821
	PRE	0.903	0.574	0.167	0.919	0.452	0.429	0.574
NaiveBayes	AUC	0.78	0.805	0.68	0.722	0.681	0.886	0.759
	SEN	0.677	0.614	0.143	0.778	0.447	0.778	0.573
	SPE	0.75	0.856	0.938	0.552	0.811	0.931	0.801
	PRE	0.798	0.794	0.926	0.799	0.704	0.955	**0.829**

3.3 Results Comparison on the Training and Test Datasets

In this challenge, participants were required to come up with an approach that generate an automatic lung-based report generation based on the volumetric CT image. For this, the organizers provided training and test datasets. Labels of the training dataset were given to the participants. However, test dataset labels were not visible to the participants. Participants were allowed to submit up to 10 runs to the system arranged by the organizers. The organizers do evaluation of the results, and ranking participants algorithms based on the results. The results of our proposed approaches on the test dataset obtained from the organizers website is depicted in Table 5. From our proposed individual classifiers, AlexNetMV performed best on the test dataset with average and min AUC of 0.757 and 0.713, respectively. From the proposed ensemble approaches, average fusion strategy outperforms all the models with mean and min AUC of 0.767 and 0.733, respectively. When best runs of each participant are compared using mean AUC on the test dataset, our result ranked 4^{th}. Detailed results of each participant algorithm on the test dataset using multiple performance metrics can be obtained at [17]. In addition, as shown in Fig.5, performance of our proposed

DL models in both the training and test datasets is nearly the same, indicating the robustness of our model. This also provides insight on how the proposed DL system will be generalized to an unknown dataset at the real test time.

Table 5. Mean and minimum AUC of the each proposed models on the test dataset.

Models	Mean AUC	Minimum AUC
AlexNetMV	0.757	0.713
AlexNetTP	0.755	0.707
VGG19MV	0.756	0.724
VGG11TP	0.731	0.722
GoogLeNet	0.427	0.36
Average	**0.767**	**0.733**
Voting	0.757	0.727
NaiveBayes	0.759	0.714

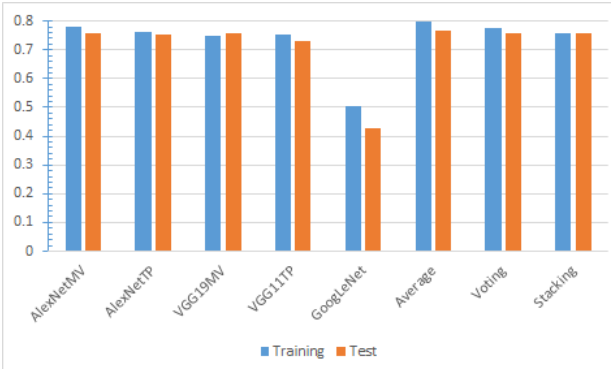

Fig. 5. Performance comparisons of the proposed models on the training and test datasets.

4 Conclusion and Future Work

In conclusion, we propose a robust CAD system for automated tuberculosis diagnosis using ensemble of different CNN architectures trained on volumetric chest CT scans of less than 300 tuberculosis patients. The proposed CNN architectures includes novel Multi-View and Triplanar-CNN architectures using pre-trained feature extraction layers of state-of-the-art deep learning models as a backend. Our experiment result that completes the top four in the challenge demonstrates that the proposed deep learning model has the ability to generate

competitive performance on automated lung-based CT report generation solely based on volumetric CT images of patients with tuberculosis.

There are still some rooms for improvement within our proposed CAD system to improve the performance. To crop the left and right lungs regions from the chest CT images, we used a fixed bounding box location for all the images through visual inspection of some random mid-level slices that could result in missing some abnormal regions of the lungs, as different CT devices produce images in different orientation. In the literature, transfer learning with different data augmentation techniques have been used to improve the performance of deep learning models in datasets with limited size. However, we only used transfer learning to increase the performance of our deep learning models on the available limited amount of training data. We did not use data augmentation techniques. Moreover, though the provided datasets were highly imbalanced, various class imbalance techniques and ensemble learner with multiple deep learning base classifiers were not investigated very well due to the limited time constraints. Ultimately, we would like to address these issues in the future.

5 Acknowledgments

This work was supported by the research fund of Çukurova University Project Number: 10683

References

1. Bhalla, A.S., Goyal, A., Guleria, R., Gupta, A.K.: Chest tuberculosis: Radiological review and imaging recommendations. Indian J. Radiol. Imaging. 25, 213–225 (2015). https://doi.org/10.4103/0971-3026.161431.
2. World Health Organization (WHO) Global Tuberculosis Report 2019, https://www.who.int/tb/global-report-2019. Last accessed 23 Jun 2020.
3. Bomanji, J.B., Gupta, N., Gulati, P., Das, C.J.: Imaging in tuberculosis. Cold Spring Harb. Perspect. Med. 5, 1–23 (2015). https://doi.org/10.1101/cshperspect.a017814.
4. Yin, J., Lu, M., Gao, L., Guo, X.: A framework of predicting drug resistance of lung tuberculosis by utilizing radiological images. In: Proceedings - 2018 10th International Conference on Advanced Computational Intelligence, ICACI 2018. pp. 308–312. Institute of Electrical and Electronics Engineers Inc. (2018). https://doi.org/doi.org/10.1109/ICACI.2018.8377474.
5. Van't Hoog, A.H., Meme, H.K., Van Deutekom, H., Mithika, A.M., Olunga, C., Onyino, F., Borgdorff, M.W.: High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. Int. J. Tuberc. Lung Dis. 15, 1308–1314 (2011). https://doi.org/10.5588/ijtld.11.0004.
6. Swanly, V.E., Selvam, L., Kumar, P.M., Renjith, J.A., Arunachalam, M., Shunmuganathan, K.L.: Smart spotting of pulmonary TB cavities using CT images. Comput. Math. Methods Med. 2013, (2013). https://doi.org/10.1155/2013/864854.
7. Xu, Z., Bagci, U., Kubler, A., Luna, B., Jain, S., Bishai, W.R., Mollura, D.J.: Computer-aided detection and quantification of cavitary tuberculosis from CT scans. Med. Phys. 40, (2013). https://doi.org/10.1118/1.4824979.

8. Harris, M., Qi, A., Jeagal, L., Torabi, N., Menzies, D., Korobitsyn, A., Pai, M., Nathavitharana, R.R., Ahmad Khan, F.: A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLoS One. 14, e0221339 (2019). https://doi.org/10.1371/journal.pone.0221339.
9. Melendez, J., Sánchez, C.I., Philipsen, R.H.H.M., Maduskar, P., Dawson, R., Theron, G., Dheda, K., Van Ginneken, B.: An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Sci. Rep. 6, (2016). https://doi.org/10.1038/srep25265.
10. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521, 436–444 (2015). https://doi.org/10.1038/nature14539.
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. pp. 2278–2323 (1998). https://doi.org/10.1109/5.726791.
12. Graziani M., Andrearczyk V., Marchand-Maillet S., Müller H.: Concept attribution: Explaining CNN decisions to physicians. Comput. Biol. Med. 103865 (2020). https://doi.org/10.1016/j.compbiomed.2020.103865.
13. Eriş, H., Çevik, U.: Implementation of Target Tracking Methods on Images Taken from Unmanned Aerial Vehicles. In: SAMI 2019 - IEEE 17th World Symposium on Applied Machine Intelligence and Informatics, Proceedings. pp. 311–316. Institute of Electrical and Electronics Engineers Inc. (2019). https://doi.org/10.1109/SAMI.2019.8782768.
14. Moon, W.K., Lee, Y.W., Ke, H.H., Lee, S.H., Huang, C.S., Chang, R.F.: Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Methods Programs Biomed. 190, 105361 (2020). https://doi.org/10.1016/j.cmpb.2020.105361.
15. Ionescu, B., Müller, H., Péteri, R., Dang-Nguyen, D.T., Zhou, L., Piras, L., Riegler, M., Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Chamberlain, J., Clark, A., Campello, A., Seco de Herrera, A.G., Ben Abacha, A., Datla, V., Hasan, S.A., Liu, J., Demner-Fushman, D., Pelka, O., Friedrich, C.M., Dicente Cid, Y., Kozlovski, S., Liauchuk, V., Kovalev, V., Berari, R., Brie, P., Fichou, D., Dogariu, M., Stefan, L.D., Constantin, M.G.: ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 533–541. Springer (2020). https://doi.org/10.1007/978-3-030-45442-5-69.
16. Arampatzis, A., Kanoulas, E., Theodora, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. In: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). vol.12260. Springer, Thessaloniki, Greece (2020).
17. 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 (2019).
18. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2020 - Automatic CT-based Report Generation. In: CLEF2020 Working Notes. http://ceur-ws.org, Thessaloniki, Greece (2020).
19. Mossa, A.A., Yibre, A.M., Çevik, U.: Multi-view CNN with MLP for diagnosing tuberculosis patients using CT scans and clinically relevant metadata. In: CEUR Workshop Proceedings. CEUR-WS (2019).

20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM. 60, 84–90 (2017). https://doi.org/10.1145/3065386.
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594.
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (2015).
23. Dicente Cid, Y., del Toro, O.A., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in CT volumes. In: Goksel, O., del Toro, O.A., Foncubierta-Rodriguez, A., and Müller, H. (eds.) Proceedings of the (VISCERAL) Anatomy Grand Challenge at the 2015 (IEEE ISBI). pp. 31–35. CEUR-WS (2015).
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z.D., Research, A.I., Lin, Z., Desmaison, A., Antiga, L., Srl, O., Lerer, A.: Automatic differentiation in PyTorch. In: Advances in Neural Information Processing Systems 32. pp. 8024–8035 (2019).
25. Liu, Y., Yao, X.: Ensemble learning via negative correlation. Neural Networks. 12, 1399–1404 (1999). https://doi.org/10.1016/S0893-6080(99)00073-8.
26. Mossa, A.A., Çevik U.: Triplanar-CNN for Automated Grading of Gliomas Using Preoperative Multi-modal MR Images. In: Proc. Of the International E-Conference onAdvances in Engineering,Technology and Management -ICETM 2020. pp. 21–27. SEEK Digital Library (2020). https://doi.org/10.15224/978-1-63248-188-7-05.