Analysis of Tuberculosis Severity Levels From CT Pulmonary Images Based on Enhanced Residual Deep Learning Architecture

Xiaohong W. Gao, Carl James-Reynolds, Ed Currie

Computer Science Department, Middlesex University, London NW4 4BT, United Kingdom

x.gao@mdx.ac.uk, c.james-reynolds@mdx.ac.uk, e.currie@mdx.ac.uk

Abstract

This research investigates the application of CT pulmonary images to the detection and characterisation of TB at five levels of severity, in order to monitor the efficacy of treatment. To contend with smaller datasets (i.e. in hundreds) and the characteristics of CT TB images in which abnormalities occupy only limited regions, a 3D block-based residual deep learning network (ResNet) coupled with injection of depth information (depth-Resnet) at each layer was implemented. Progress in evaluation has been accomplished in two ways. One is to assess the proposed depth-Resnet in prediction of severity scores and another is to analyse the probability of high severity of TB. For the former, delivered results are of 92.70 ± 5.97% and 67.15±1.69% for proposed depth-Resnet and ResNet-50 respectively. For the latter, two additional measures are put forward, which are calculated using (1) the overall severity (1 to 5) probability, and (2) separate probabilities of both high severity (scores of 1 to 3) and low severity (scores of 4 and 5) respectively, when scores of 1 to 5 are mapped into initial probabilities of (0.9, 0.7, 0.5, 0.3, 0.2) respectively. As a result, these measures achieve the averaged accuracies of 75.88% and 85.29% for both methods respectively.

Keywords: Deep learning, residual deep learning network, classification, 3D block-based image classification, Tuberculosis (TB), severity score of TB.

1. Introduction

While deep learning (DL) led networks have achieved cutting edge performance in many fields, they rely on the availability of large amounts of data as training sets. In many fields, this prerequisite is difficult to meet, especially in medically related research. One way to respond to this challenge is data augmentation, to process data at various levels and to concatenate information acquired at multiple scales (through multiple convolutions) at each layer. These networks (for example, Inception-ResNet [1]), are able to maintain the computational cost constant while accelerating the training process considerably. This is achieved through the introduction of batch normalisation (BN) (ranging values within [0,1]), a reduction/projection layer (by introducing a 1×1 convolution filter) and by modelling a residual unit instead of stacking layers on top of each other. This paper investigates the feasibility of applying the inception-Resnet architecture to analysing the level of severity of tuberculosis (TB) from 3D pulmonary CT images. This is one of the challenges facing TB diagnosis where hand-crafted methods, e.g. texture-based approaches, lack generality.

Tuberculosis (TB) constitutes one of the top 10 causes of death worldwide, but can be cured using appropriate treatment with a course of antibiotics if diagnosed in a timely fashion. Conventional diagnostic procedures involve growing a microbiological culture; a process that is expensive in terms of cost and time. For this reason, high resolution computer tomography (CT) of pulmonary (lung) images has been used to aid clinicians in expediting diagnosis and for monitoring prognosis when administering antibiotic drugs. The infectious disease TB is caused by exposure to Mycobacterium Tuberculosis (M. TB) through the inhalation of tiny droplets from the coughs or sneezes of an infected person and remains one of the top 10 causes of death worldwide. In 2015, 10.4 million people fell ill with TB, of whom 1.8 million died of the disease [2]; 0.4 million of these were HIV patients. While most TB cases occur in developing countries**,** this disease, prevalent in the Victorian era, has still not been eradicated in developed countries. On the contrary, the rate of the disease has recently risen in some areas of western countries; for example, in London UK, for various reasons, including drug abuse and homelessness.

Although TB remains a serious contagious condition, it can be cured if promptly treated with suitable antibiotics. For varying degrees of TB severity, different amounts and combinations of antibiotics will need to be administered to treat the disease. To clinically detect the level of TB severity, the most definitive method is to grow a microbiological culture; an expensive procedure, that can take several months. Therefore, there is an urgent clinical need for additional methods that can determine TB severity quickly, accurately and economically. One approach is to apply non-invasive high resolution *Computer Tomography* (CT) imaging to assist clinicians in analyzing, diagnosing and delivering optimal treatment for TB patients.

This paper focuses on the application of state of the art deep learning techniques to the analysis of CT pulmonary

images and is organised as follows. Section 2 reviews both the diagnostic procedures for detecting the level of severity of TB and existing deep learning techniques. In Section 3, the datasets and proposed methodology to score the severity of TB are described. In Section 4, the implementation details are specified, together with experimental results. Section 5 summarises the research conducted, discusses its limitations and indicates future directions.

2. Background

2.1 Tuberculosis diagnosis based on CT lung images

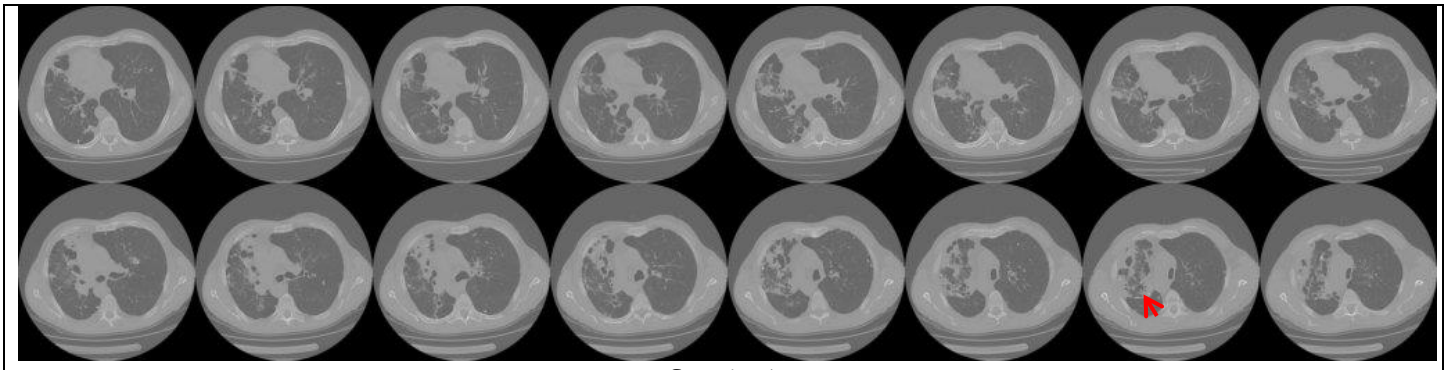
Mycobacterium tuberculosis (M. TB) was discovered 130 years ago and is an aerobic, non-motile, non-spore-forming rod bacterium that is highly resistant to drying, acid, and alcohol. This bacterium transmits from person to person via droplet nuclei containing the airborne organism, mainly by coughing. A person with active but untreated TB is estimated to go on to infect an average of 10 to 15 other people per year, depending on the number of droplets expelled by the carrier, the duration of exposure, and the virulence of the M. TB [3].

Clinically, the definitive diagnosis of active tuberculosis is the detection of the presence of the M.TB bacterium, the causative microorganism of TB, which can be achieved through growing a microbiological culture from tissue taken from the patient [4]. In practice, however, the culture growth of M. TB usually takes 2 or more weeks. Hence to expedite diagnosis of active TB, an array of combined approaches are employed, including a tuberculin skin test (TST), blood test, amplification of M. TB nucleic acids and/or pathological examinations from biological specimens. While these methods assist diagnosis to a large degree, they are not specific in determining severity.

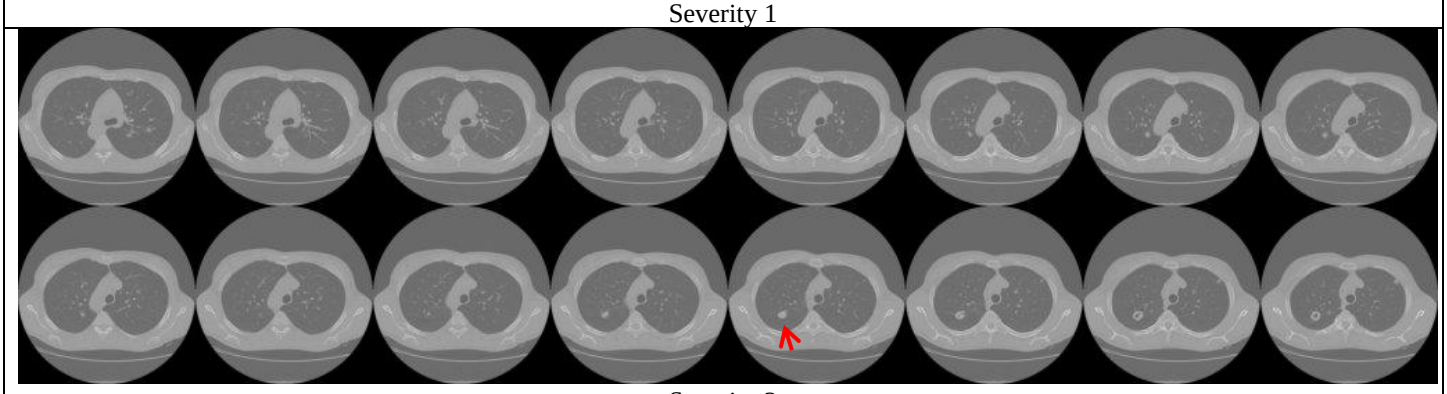
Since pulmonary TB presents characteristic patterns in the lung, radiological imaging is an invaluable tool to assist diagnosis, including chest X-rays and CT. Conventional chest X-rays remain the most commonly employed method for screening, diagnosis and the follow up of treatment responses in patients with pulmonary TB. However, high-resolution CT of the chest appears to be more sensitive than X-rays in identifying early parenchymal lesions, detecting mediastinal lymph node enlargements and determining disease activity in TB [5-7].

Clinically, diagnosis of TB is based on the observation of a number of factors that contribute to the so-called ‘index of suspicion’ [6]. If TB is detected in a timely manner and fully treated, people with the disease can quickly become noninfectious and eventually cured. Therefore, early diagnosis and treatment are crucial for both maintaining patients’ health and reducing the proliferation of TB to the public. Figure 1 demonstrates five levels of TB severity, scored from 1 to 5, with 1 referring to the most severe and 5 the least severe manifestation of the

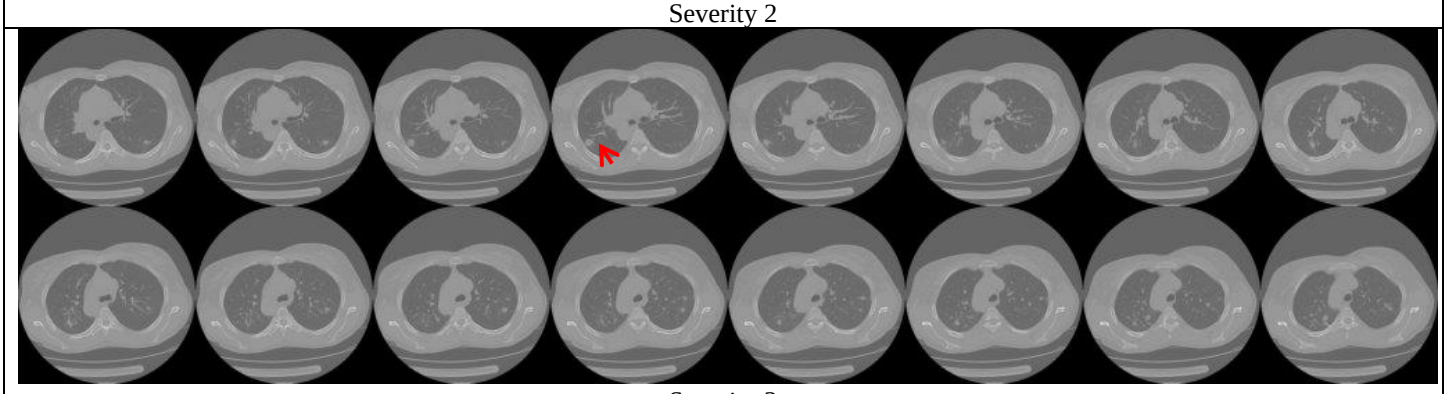
disease. For each 3D CT volume, only the middle 16 slices are presented, in which arrows point to the infected regions. As can be seen from Figure 1, many of the slices from different severity categories present similar patterns, making the classification of severity a challenging task.



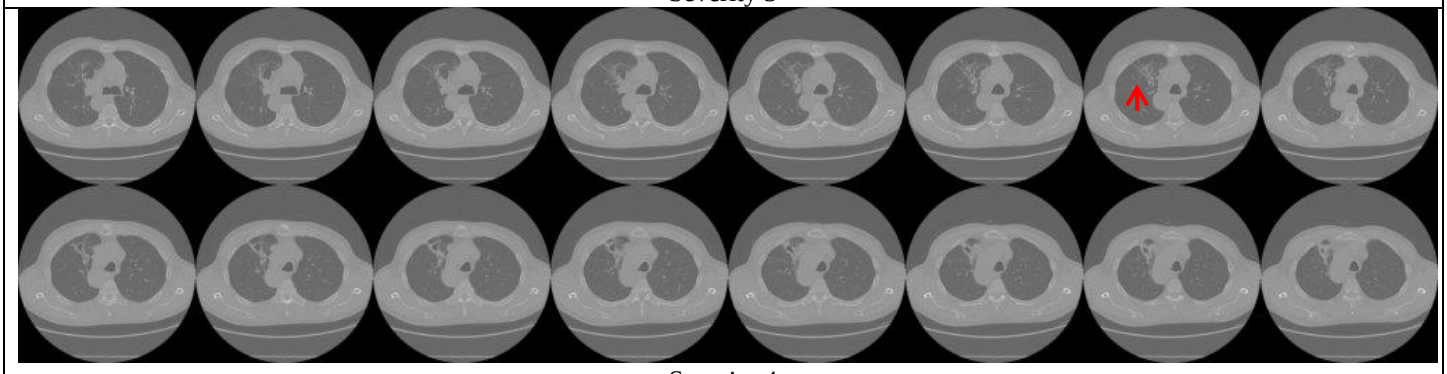
Severity 1



Severity 2



Severity 3



Severity 4

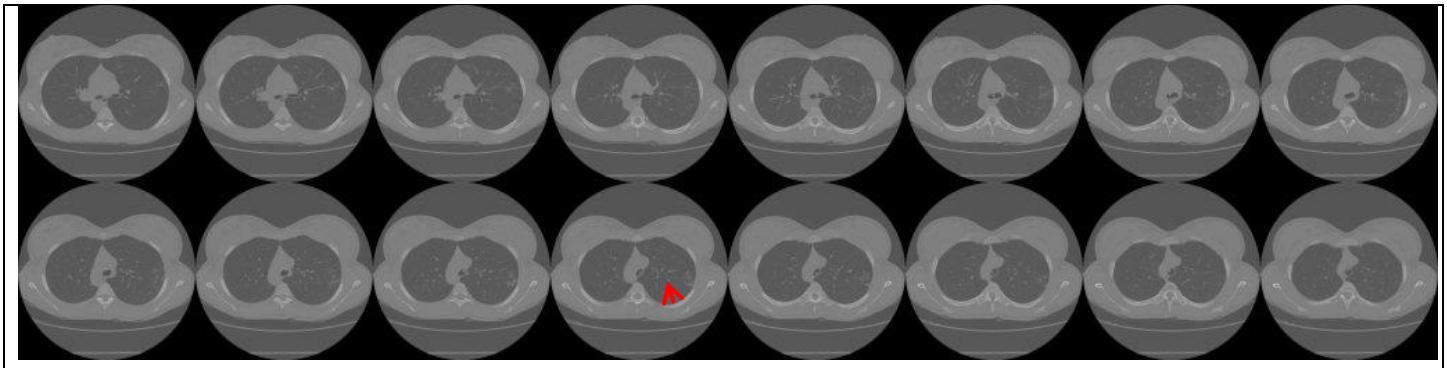

Severity 5

Figure 1. Five levels of severity of TB disease scored 1 (top) to 5 (bottom) presented using CT lung images where arrows point to the infected regions. Only the 16 middle slices from each 3D volume are shown.

2.2. Deep learning in medical applications

Deep learning neural networks refer to a class of computing machines that can learn a hierarchy of features by establishing high-level features from low-level ones and was pioneered by Fukushima [8] based on biologically-inspired human vision systems. One of these models is the convolutional neural network (CNN) developed by LeCun et al. [9]. Consisting of a set of machine learning algorithms, CNN is comprised of several (deep) layers of processing involving learnable operators (both linear and non-linear), and hence has the ability to learn a hierarchy of information by building high-level information from low-level ones, thereby automating the process of construction of discriminative information [10]. In addition, recent advances in computer hardware technology (e.g. the *Graphics Processing Unit* (GPU)) have propagated the implementation of CNNs in representing images.

Conventionally, training a DL model requires large datasets and substantial training time. For example, the pre-trained CNN classifier, Alexnet [11], is built on 7 layers, simulating 659K neurons with 60 million (M) parameters and 630M connections, and trained on a subset (1.2M with 1K categories) of ImageNet [12] with 15M 2D images of 22K categories, taking up 16 days on a CPU and 1.6 days on a GPU.

DL-oriented approaches are widely applied to large quantities (often millions) of images; they have recently been applied to medical images in a range of domains and achieved state of the art results. In particular, CNN- based approaches have won a number of competitions, including the Kaggle competition on detection of diabetic retinopathy [13] and segmentation of brain tumours from MRI images [14].

In the medical domain, not only are the number of datasets limited (usually in hundreds), but also images are in multiple dimensions, ranging from 2D to 5D (e.g. a moving heart at a specific location). Hence additional measures have to be taken into account in order apply DL techniques. For example, to classify 3D echocardiography video images, Gao et al [15] designed a fused CNN architecture to incorporate both

unsupervised CNN and hand crafted features to leverage the shortage of datasets. In addition, to capitalize on the information that a medical image proffers, they integrated two networks that were implemented for 2D and 3D respectively, for classification of CT brain images [16]. With regard to TB data, where only small regions of each slice and only a few slices in a volume present infected disease, one way to increase the amount of datasets is to divide each slice into smaller segments or patches as implemented in [17, 18] applying a patch-based deep learning technique to analysis of TB images for classification of TB types and analysis of multiple drug resistance.

Theoretically, a CNN can be conveyed as a process of minimising a cost function between ground truths and predictions. To this end, with a set of training data $\left(x^{(i)}, y^{(i)}\right)$, where image $x^{(i)}$ is in three-dimension (inclusive of RGB channel as the 3rd dimension. Note: DL is a general approach and treats any input image as colour data with dimensions of $(row, column, RGB) = (M, N, 3)$ whereby $(M, N, 1)$ is red, $(M, N, 2)$ is green and $(M, N, 3)$ is blue. For a grey image, 2D representation using (M, N) should be sufficient) and $y^{(i)}$ the indicator vector of affiliated class of $x^{(i)}$, i.e. the ground truth, a CNN network is used to solve the equation expressed in Eq. (1). In doing so, the feature maps of an image, namely, $w_1, \dots, w_L$, will be learnt, a process known as deep learning.

$$\underset{w_1,\dots,w_L}{argmin} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(f\left(x^i; w_1, \dots, w_L\right), y^i\right) \tag{1}$$

where $\mathcal{L}$ refers to a suitable loss function (e.g. the hinge or log loss) and f the selected classifier.

As a result, a CNN architecture can be constructed by stacking multiple layers of convolution and subsampling in an alternating fashion. While a CNN network can be enhanced into going deeper by piling a large number of layers, sometimes the increased depth appears to have little contribution to the accuracy of a trained model. This is due to the well-known vanishing gradient obstacle, i.e. as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitely small. Consequently, as the network becomes deeper, its performance gets saturated or even starts degrading rapidly. Although several remedy strategies have been reported to tackle the vanishing gradient barrier (for instance, adding an auxiliary loss [19] in a middle layer as an extra supervision), none seem to really address the problem thoroughly.

Recently, deep residual networks (ResNet) [20, 21] introduce the notion of '*identity shortcut connection*' that bypasses one or more layers as illustrated in Figure 2, which demonstrates a residual block where ReLU refers to a rectified linear unit to ensure the data are greater than zero (>0) and batch normalisation is used to convert all the matrix elements to values between [0,1] in order to speed up the calculation. A key advantage of residual units is that their skip connections allow direct signal propagation from the first to the last layer of the network,

especially during backpropagation. This is due to the fact that gradients are propagated directly from the loss layer to any previous layer while skipping intermediate weight layers that have potential to trigger vanishing or deterioration of the gradient signal.

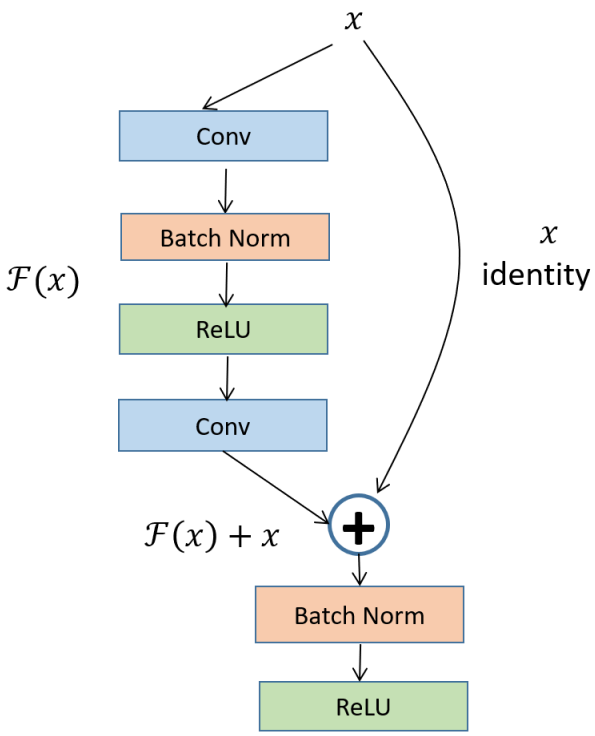


Figure 2. A typical residual block in a residual learning network Resnet [20] and its optimised version.

If $\mathcal{H}(x)$ refers to an underlying mapping to be fitted by a few stacked layers (not necessarily the entire net) with x denoting the inputs to the first of these layers, then the network that approximates complicated functions can equivalently approximate the residual functions $\mathcal{H}(x) - x$, assuming both inputs and outputs are of the same dimensions. Therefore these layers can be applied to approximate a residual function in Eq. (2).

$$\mathcal{F}(x) \coloneqq \mathcal{H}(x) - x \qquad (2)$$

which leads to Eq. (3).

$$\mathcal{H}(x) \coloneqq \mathcal{F}(x) + x \qquad (3)$$

As shown in Figure 2, a building block to every few stacked layers is defined in Eq. (4) where the operation $\mathcal{F} + x$ is performed by a shortcut connection and element-wise addition.

$$y_l = \mathcal{F}(x_l, \{W_i\}) + x_l \qquad (4)$$

where x_l and y_l are the input and output vectors at layer l and the dimensions of x_l and $\mathcal{F}$ must be equal. Otherwise, a linear project W_s can be performed by the shortcut connection to match these two dimensions as formulated in Eq. (5).

$$y_l = \mathcal{F}(x_l, \{W_i\}) + W_s(x_l) \qquad (5)$$

While the training mainly focuses on deep residual learning, stacking layers should not degrade the network performance. This is because those layers do not do anything in relation to vanishing gradients apart from simply stacking identity mappings upon the current network, hence resulting in similar architecture performance. This indicates that a deeper model should not produce a training error higher than its shallower counterpart.

2.3 Inception deep convolutional architecture

While a ResNet can perform in a deeper manner, an inception network can go wider by inserting extra components at each layer. The inception deep convolutional architecture was introduced in [19] firstly as Inception-v1. Later the inception architecture was refined in various ways in order to accelerate the training time and reduce the computation cost, firstly by the introduction of batch normalization [22] (Inception-v2) to normalise all values within the range of (0,1), then by the addition of factorization [23] (Inception-v3) to transform a large convolution (e.g. 5×5) into two smaller ones (e.g. 3×3 and 3×3).

A typical residual-inception network follows a split-transform-merge paradigm. Each inception block begins with a filter reduction layer (1×1×1), then performs convolution layer (e.g. 3×3×1), and finally completes with a filter-expansion convolution layer (1×1×1) without activation, the layer that is used for scaling up the dimensionality of the filter banks (or maps) before the addition to match the depth of the input. This is needed to compensate for the dimensionality reduction induced by the inception block.

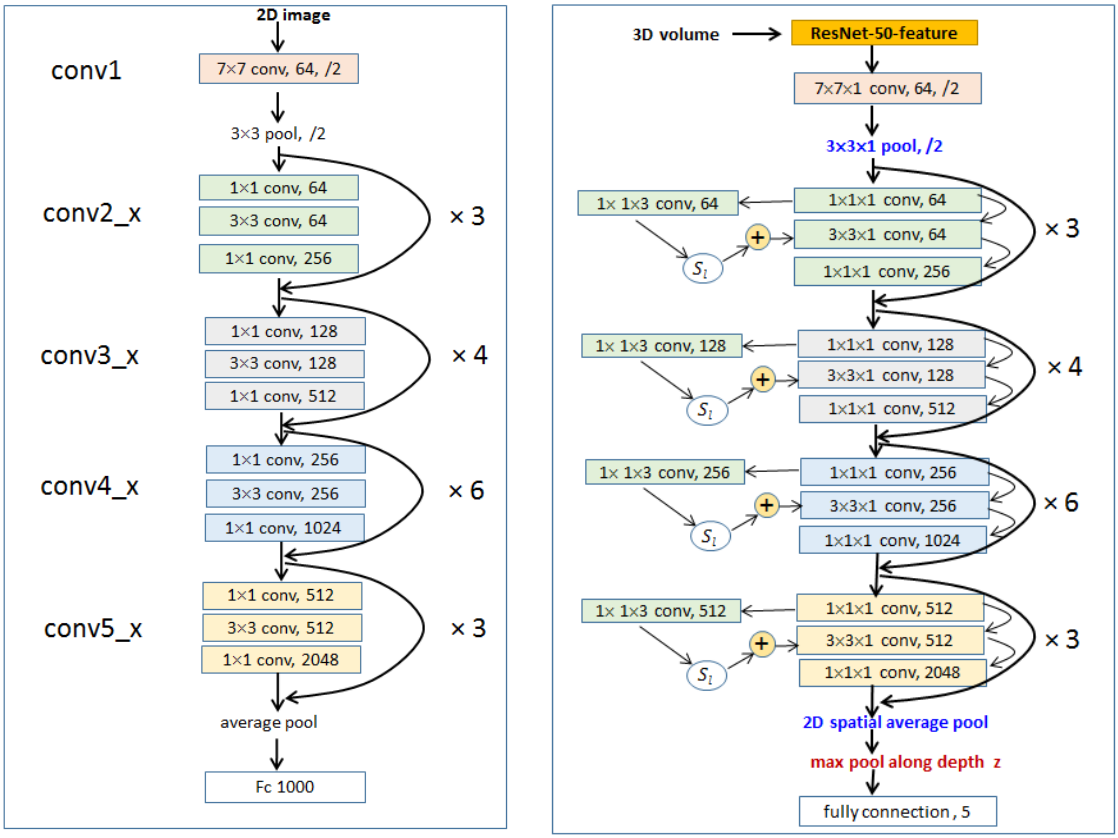
In this work, an enhanced inception-Resnet, i.e. depth-Resnet is applied for analysis of the level of severity of tuberculosis from CT lung images.

3. Methodology

3.1 Depth-Resnet

Inspired by the temporal residual network [24, 25] and inception-Resnet architecture, the network is built on the ResNet-50 model and is illustrated in Figure 3, where the left-hand graph is the original model and the right-hand graph the enhanced architecture depth-Resnet, applied in this study. The information along the third direction (z)

for 3D TB datasets is embedded with 3 layers in each block as further depicted in Figure 4. As illustrated in Figure 3(a), the architecture of ResNet incorporates 50 layers whereas the model, ResNet-50, is trained on ImageNet data with 1000 classes. A typical block comprises of 3 convolutional operations, consisting of 1×1 dimensionality reduction, 3×3 spatial aggregation and 1×1 dimensionality restoration of filtering operations, in addition to normalisation and ReLU layers when addressing two-dimensional images. In this study, depth-Resnet is built on the pre-trained ResNet-50 model by replacing the last classification (prediction) layer, which is followed by the enhanced convolutional layers of conv2_x to conv5_x as elaborated in Figure 3(b) to incorporate depth information.



(a) Original ResNet-50 model (b) Depth-ResNet

Figure 3. The original ResNet-50 model (a) and the Inception-ResNet-50 (b) architecture applied in this paper, where ×N at each conv level refers to the block (e.g. conv5_x) repeats N (e.g. 3) times consecutively.

To take advantage of ResNet-50 using 3x3 filters to perform spatial convolution, the depth convolution also adopts 3 pixels, i.e. 1x1x3 between the current, front and back frames. Since some 3D blocks only contain 10 frames, the chosen block size is 8 frames. Therefore the interval (stride) between the current frame and the front (or the back) is set to be between 1 and 7, to be selected randomly. In the end, to minimise the classification errors, a global pooling layer followed by a 5-way fully connected layer, optimised using a Softmax approach is conducted.

In Figure 3(a), for each residual unit, the input feature map $x_l \in \mathbb{R}^{H\times W\times D\times C}$, where H, W, D are the spatial dimensions along the height, width, and depth directions for a 2D dataset and C the feature dimension. Such maps can be thought of as stacking 2D spatial maps of C dimensional features along the depth (z) dimension. At layer l with input x_l, a residual block is defined as Eq. (6).

$$x_{l+1} = f(x_l + \mathcal{F}(x_l; \mathcal{W}_l)) \tag{6}$$

where $f \equiv ReLU$, $\mathcal{W}_l = \{W_{l,k} | 1 \leq k \leq \mathcal{K}\}$, and $\mathcal{K} = 3$, with $\mathcal{F}$ denoting the residual function representing the convolutional operations by convolutional filter weights $\mathcal{W}_l$.

Each of the $\mathcal{K}$ layers in the l^{th} residual unit performs the filtering operation as formulated in Eq. (7).

$$x_{l,k+1} = W_{l,k} x_{l,k} \tag{7}$$

where $W_{l,k} | 1 \leq k \leq \mathcal{K}$ are the convolutional filter kernels arranged as a matrix. For simplicity, batch normalisation layers as shown in Figure 2 are omitted in Figure 3. Hence, the residual unit is expressed in Eq. (8).

$$\mathcal{F} = f\left(W_{l,3} f\left(W_{l,2} f\left(W_{l,1} x_l\right)\right)\right) \tag{8}$$

On the other hand, in Figure 3(b) of depth-Resnet, built on the inception concept, the depth convolution block operates on the dimensionality reduced input, $x_{l,z}$ with a bank of 3D filters, $W_{l,z}$. Biases $b \in \mathbb{R}^C$ are also applied with initial values of 0 as formulated in Eq. (9).

$$x_{l,z} = W_{l,z} x_{l,1} + b \tag{9}$$

As a result, Eq. (8) in Figure 3(a) becomes Eq. (10) in Figure 3(b).

$$\mathcal{F} = f\left(W_{l,3}\left(S_z f\left(x_{l,z}\right) + f\left(W_{l,2} f\left(W_{l,1} x_{l,1}\right)\right)\right)\right) \tag{10}$$

where S_l is affine scaling along depth direction with a bias between 0 and 0.01. This scaling is adaptive to facilitate generalisation performance and will be learnt during the training of the network. Figure 4 elaborates the depth receptive field of a single neuron. In Figure 4, the convolution at each convolution layer along the depth (z) direction ($x_{l,z}$) takes place between 3 neighbouring slices or feature maps, i.e. front, current, and back, with randomly chosen stride (between 1 and 7 in this study). This feature is then added to the block with a scaling factor as a component of the residual unit. The pooling involves two stages. The *avg-pool* occurs for 2D spatial

global average pooling whereas *max-pool* is conducted along z direction performing global max pooling upon those feature maps.

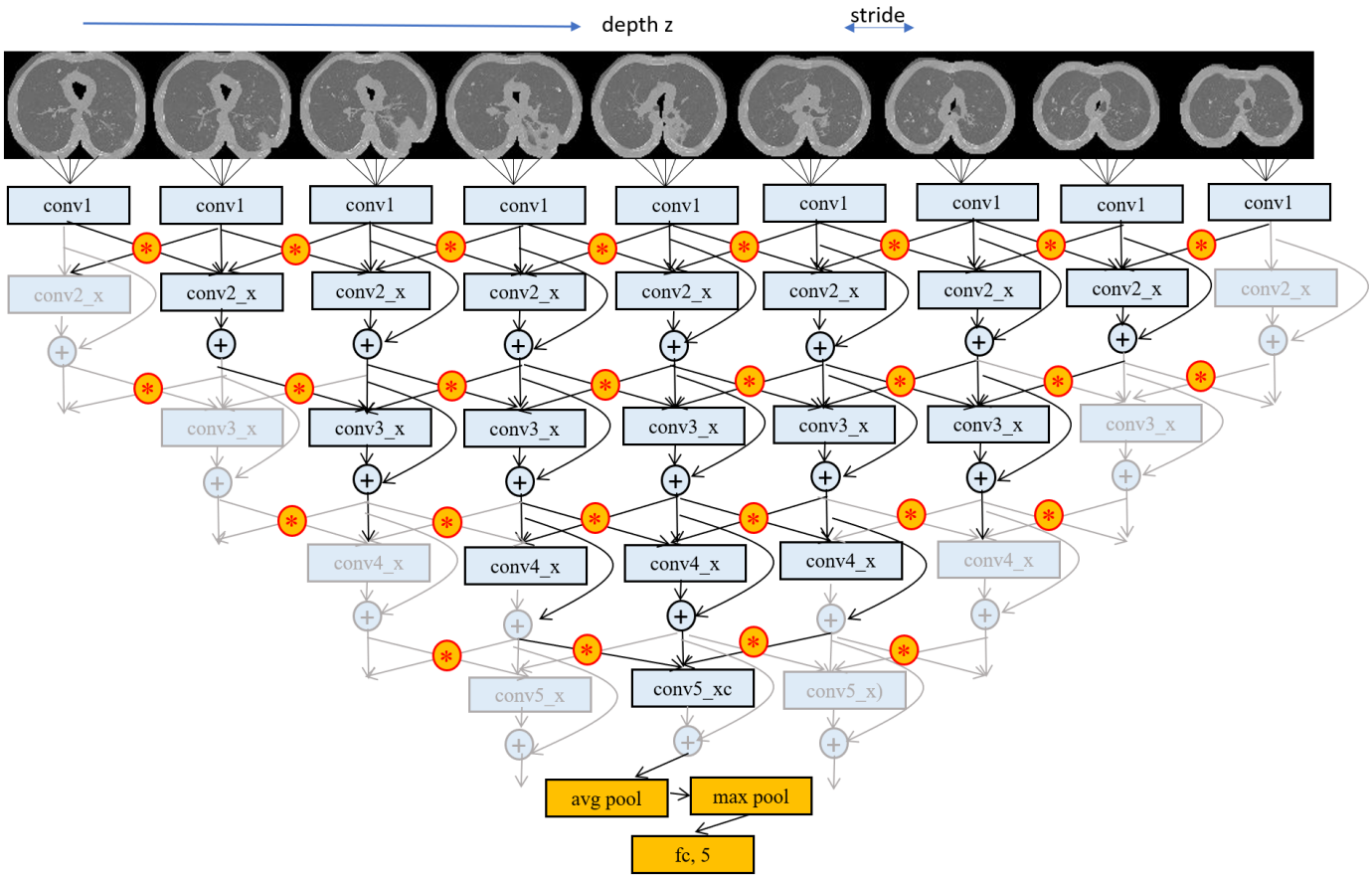


Figure 4. A block in the depth-Resnet that is applied in the paper. The outputs of conv5_X are max-pooled in time and fed to the fully connected (fc) layer of the proposed depth-Resnet as shown in Figure 3(b) to classify 5 categories.

On the other hand, to integrate block scores into a volumetric label for each dataset, a support vector machine (SVM) [27] is applied. To train a SVM classifier, linear optimisation is applied to minimise formula Eq. (11).

$$L_w = \left[\frac{1}{n}\sum_{i=1}^{n} max\left(0,1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2 \qquad (11)$$

where $x_i$and y_i refer to input and ground truth respectively, with w the weight and incept b to be trained.

SVMs are a set of supervised learning models that analyse and classify data applying associated learning algorithms. There are linear and non-linear SVMs. While in a linear SVM, such as the one employed in this study, any hyperplane can be written as the set of points x satisfying Eq. (12).

$$w \cdot x - b = 0 \qquad (12)$$

3.2 Datasets

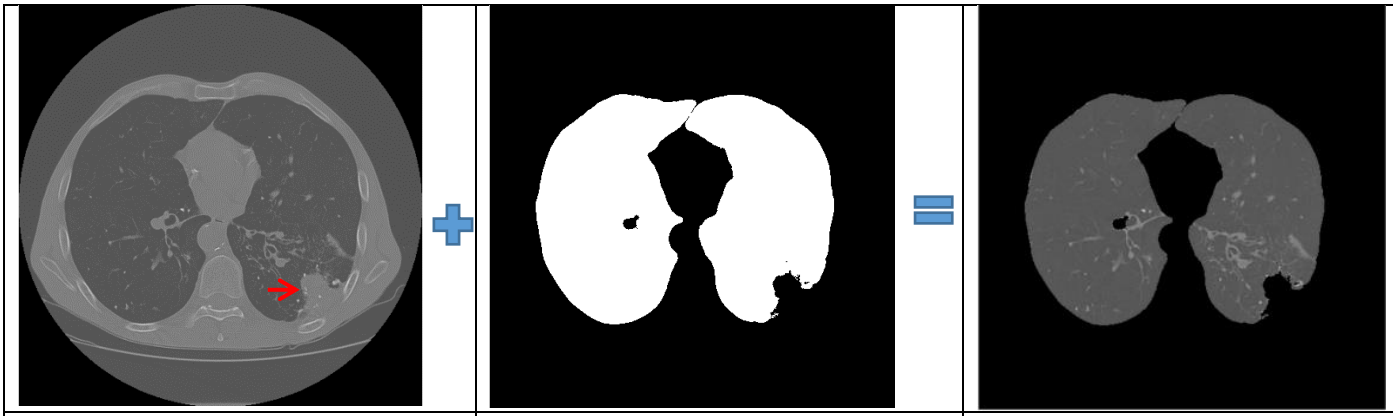
Data from the competition organised by ImageCLEF2018 on Tuberculosis severity scoring task (task#3) [28, 29] was used, including chest CT scans of TB from 170 patients with the corresponding severity scores (1 to 5) and the severity levels designated as "*high*" and "*low*", which contains 90 low severity (with scores 4 and 5) and 80 high severity (with scores 1, 2 and 3) as listed in Table 1. Each volume of the dataset has a 2D spatial dimension of 512×512 pixels per slice and varying number of slices (between 50 and 400) along depth (z) direction.

Table 1. The number of datasets and blocks applied for both training and testing with corresponding severity scores.

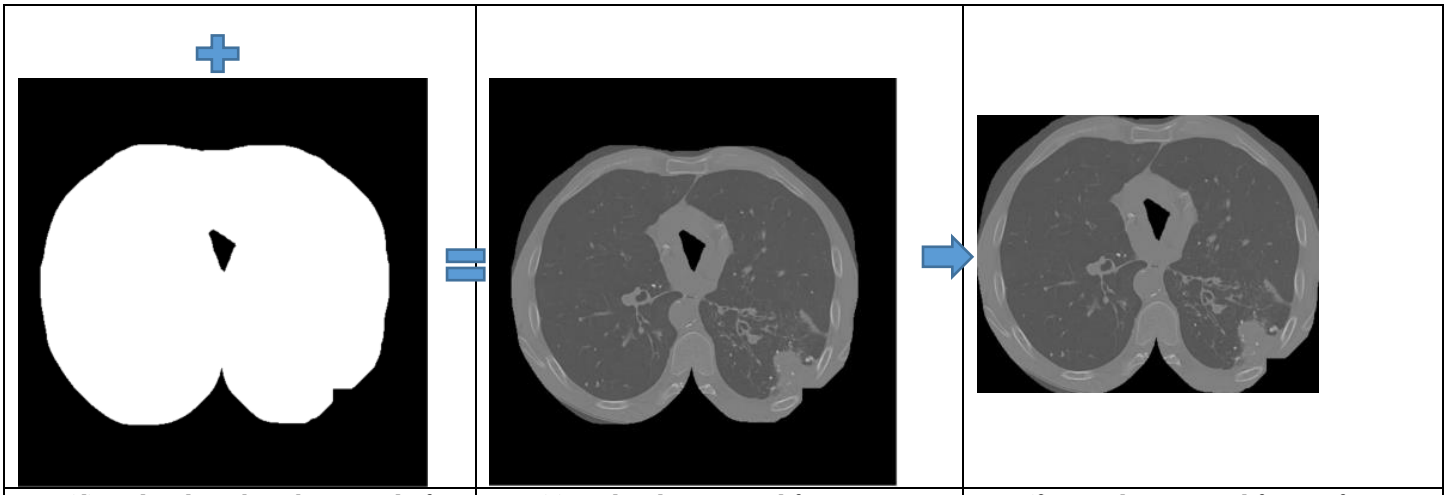
	High Severity			**Low Severity**		**Total**
Severity	1	2	3	4	5	
Train/evaluation	10	10	30	40	10	100
Train Blocks	978	1108	2976	3869	967	9898
Test	5	7	18	33	7	70
Test Blocks	496	801	1777	3265	676	7015
Total	15	17	48	73	17	**170**
	1,474	1,909	4,753	7,134	1,643	**16,913**

3.3 Image data pre-processing

Before the training of severity models, volumes of images firstly undergo the segmentation process to remove surrounding artefacts. Since the masks that are supplied with the images from ImageCLEF sometimes over remove lung boundaries, dilation of masks are performed first. The balance is struck for this collection by dilating using 30 pixels to ensure that not too much of the unintended boundaries are included. Figure 5 illustrates this dilation process, where the top row presents the segmentation result (c) applying the original mask (b) for slice (a) and the bottom row depicts the segmented slice (e) with dilated mask (d). Figure 5(f) shows the final segment after removing the background from segment (e). The arrow in (a) points to the abnormal region of interest, which is missing in segment (c).



(a) Original sliice of a volumetric image | (b) Corresponding mask | (c) Segmented image



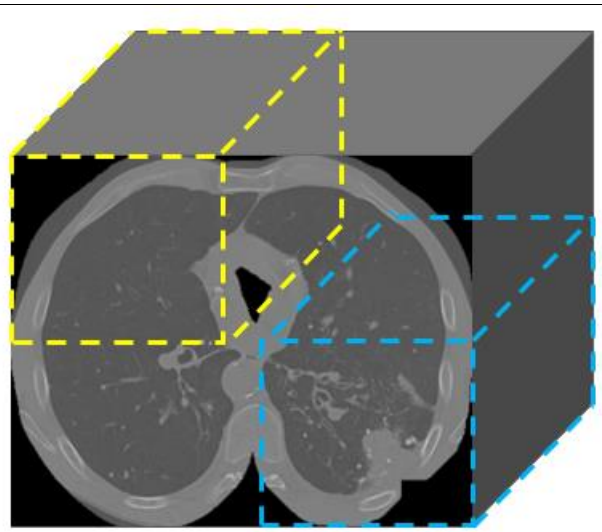
(d) Dilated mask with 30 pixel of disk

(e) Dilated segmented frame

(f) Final segmented frame after removing background of (e).

Figure 5. The process of segmentation with dilated masks. Top row: the segmented slice (c) with the original mask (b). Bottom row: dilated mask (d) from (b) applied to the original slice (a) producing (e) and finally (f) after removing background of (e).

Then, upon the segmented volume of $460 \times 340 \times z$, 24 blocks of size of $128 \times 128 \times z$ are created with an overlap of ~64 pixels as illustrated in Figure 6.



(a)

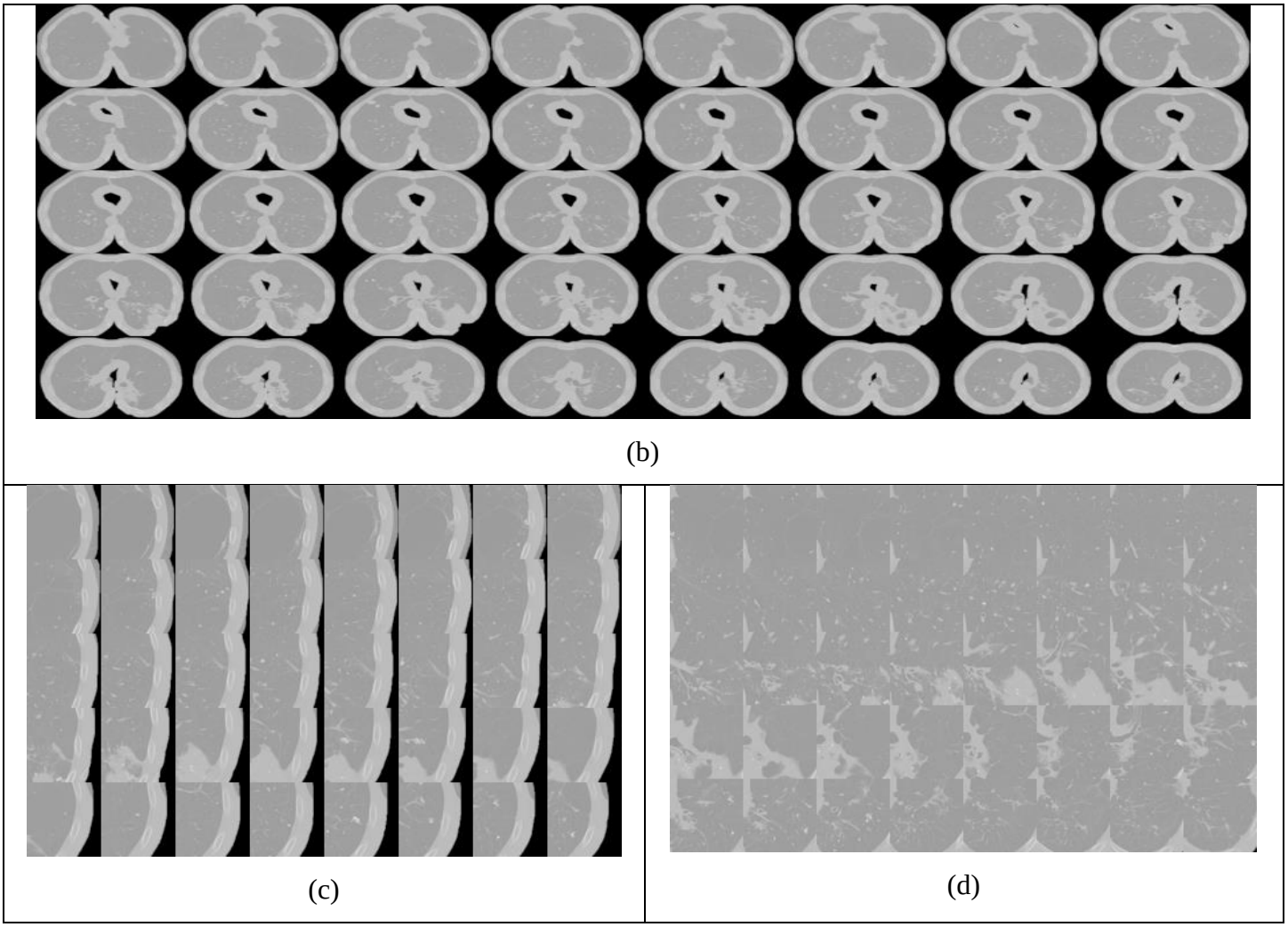


Figure 6. Illustration of segmented 3D volume (a); its montage with equally spaced selected frames (b); and two of its blocks presented using montage style with equally spaced selected frames (c) and (d).

Since some corner blocks comprise large amount of background information, i.e. pixel value is 0, these frames, in particular at front and back of a volume along the z direction, are removed when the background occupies more than one third of the total space. Hence the depth (z) of each block varies between 11 and 250 for all datasets after segmentation. As a result, many 3D volume datasets have less than 24 blocks after pre-processing. Each block has been resized to $256 \times 256 \times z$ from $128 \times 128 \times z$ to save training time.

4. Results

The training system is implemented using Matlab software built on the MatConvNet [26] toolbox, by following standard ConvNet training procedures [10, 11]. The system starts with the application of the ResNet-50 model as demonstrated in Figure 3(a) to compensate for the shortage of datasets. Then, by replacing the last prediction layer, every first and third residual unit are transformed at each convolution stage (*conv2.x* to *conv5.x* in Figure 3) with the proposed 3D residual units of Eq. (10) (Figure 3(b)). The depth filters are of dimension

$W' \times H' \times D' \times C \times C = 1 \times 1 \times 3 \times C \times C$ and are initialised to randomly selected values. The 8-slice sub-blocks are applied in this work as an input and global max pooling along depth z direction as formulated in Eq. (13) and illustrated in Figure 3(b) is conducted immediately after the 2D spatial global average pooling layer.

$$x(i,j,c) = \max_{1 \le k' \le D'} x(i,j,k',c) \tag{13}$$

The input size is of dimension $256 \times 256 \times z$ (224 with 32 pixels as borders), with z varying between 11 and 250 slices in this collection. During the training, a batch of 16 sub-blocks (128 slices in total), each containing 8 slices, was chosen from five levels of severity classes. The stride for the 8 slices in each block was randomly selected between 1 and 7.

At testing time, each dataset undertakes the same pre-processing procedure (Section 3.3) to generate 128×128×depth blocks as elaborated in Figure 6. Then the trained depth-Resnet model (Figure 3(b)) takes each block as a whole, selects 8 slices at equal depth space and propagates these slices through the trained model to produce a single prediction for this block with severity scores labeled between 1 and 5. The scoring strategy adopts a faster fully convolutional testing strategy [30, 31], which is applied to the original slices and their horizontal flips and averages the predictions from all 2D spatial locations. Subsequently, the inference can be performed in a single forward pass for the whole block. The training takes place on a Dell Precision T7600 computer with a 64-bit Ubuntu operating system and one GPU with 64 GB memory. It takes 4 days to train 100 datasets and 2 days to test 70 volumes.

Since each volume of the 3D dataset contains around 24 blocks with individual severity scores, the overall score for each patient's dataset has to be integrated from the individual block scores. In principle, the five levels of severity can be treated as 2 classes labeled as 'high' (with scores 1, 2 and 3) and 'low' (with scores 4 and 5). Hence, three measures can then be formulated to convey the inter-relationships between blocks scored 1 to 3, 4 to 5 and 1 to 5 respectively and are calculated in Eqs. (14), (15) and (16) respectively where levels of severity of 1 to 5 are assigned linearly to probabilities of 0.9, 0.7, 0.5, 0.3, and 0.1 respectively.

$$probability_{high} = \frac{0.9 \times num_{block_1} + 0.7 \times num_{block_2} + 0.5 \times num_{block_3}}{num_{block_1} + num_{block_2} + num_{block_3}} \tag{14}$$

$$probability_{low} = \frac{0.3 \times num_{block_4} + 0.1 \times num_{block_5}}{num_{block_4} + num_{block_5}} \tag{15}$$

$$probability_{all} = \frac{0.9 \times num_{block_1} + 0.7 \times num_{block_2} + 0.5 \times num_{block_3} + 0.3 \times num_{block_4} + 0.1 \times num_{block_5}}{num_{block_1} + num_{block_2} + num_{block_3} + num_{block_4} + num_{block_5}} \tag{16}$$

Hence the probability of a whole volume dataset can then be decided by these measures, which is in turn utilized to score the severity. For example, in this study, if a dataset has $probability_{high} > 0.7$ and $probability_{low} <$

0.20 and $Num_{block1} > 0$, then this dataset is classified as severity 1. In Table 2, two calculations are applied. One is based on the overall probability (Level-1) as formulated in Eq. (16), which is simple and straightforward. The Level 1 calculation was applied by the authors of this paper to the imageCLEF Tuberculosis 2018 competition [29], with a result that ranked 14th (out of 36 submissions) in terms of accuracy (AUC=0.6534), obtained using a different set of test data (n=109) with unknown severity levels.

A drawback of Level-1 is that the calculation treats all severity blocks equally, whereas higher severity volumes usually contain low severity blocks. Therefore, blocks with low severity scores 4 and 5 contribute more to the equation when it comes to the overall classification. In additional to probability range distribution, a number of other measures are also factored in, to address special cases. For example, if a dataset has 5 blocks scored 1 and another 5 scored 5, the final score based on Eqs. (14) and (15) is 3. However, there is not any individual block with a score of 3 for this volume (Level-2 measure predicts score 1). In this study, the *Level-2* calculation is applied, utilizing both $probability_{high}$ and $probability_{low}$, which delivers much improved performance. Again, the integration of two levels of probability is conducted based on a SVM. In Table 2, the results are the average from three runs (each run of training takes about 4 days) with standard deviations. During each run, the number of training sets (n=100) as listed in Table 1 is randomly selected manually and the remaining data (n=70) kept for the final test. The averaged accuracy from the Level-2 calculation is 85.29%, a nearly 10% increase from Level-1 with 75.88% accuracy. Significantly, the Level-2 calculation appears to capture high severities with scores 1 and 2 much better than Level-1 with 86% and 70% for Level-2 and 80% and 60% for Level-1 respectively.

Table 2. The accuracy performance from both Level-1 and Level-2 calculations.

Severity	**1**	**2**	**3**	**4**	**5**	**Average**
Level-1	0.80 ± 0.00	0.60 ± 0.05	0.75 ± 0.00	0.88 ± 0.02	0.82 ± 0.12	**75.88 ± 3.80%**
Level-2	0.86 ± 0.08	0.70 ± 0.01	0.77 ± 0.02	0.90 ± 0.00	0.84 ± 0.04	**85.29 ± 3.00%**

Table 3 illustrates the sensitivity and specificity values for both measures. Statistically, *sensitivity* measures the proportion of actual positives that are correctly identified as such, whereas *specificity* measures the proportion of actual negatives that are correctly identified as such. Hence high *sensitivity* and high *specificity* implies greater credibility of the classification results. The results in Table 3 are based on the concatenation of three run results; for example, the test sample size for Severity 1 was five in each individual run and is now 15 (= 3 × 5) for 3 runs. Again, Level-2 performs better than Level-1 with higher average sensitivity (84.16%) and higher specificity (95.35%) whereas Level-1 delivers 77.17% and 93.35% respectively. Overall both calculations present higher specificity, suggesting the conversions between severity level and probability as well as Eqs. (14) to (16) are sufficient to separate in-between severities.

Table 3. Comparison of sensitivity and specificity of both Level-1 and Level-2 calculations.

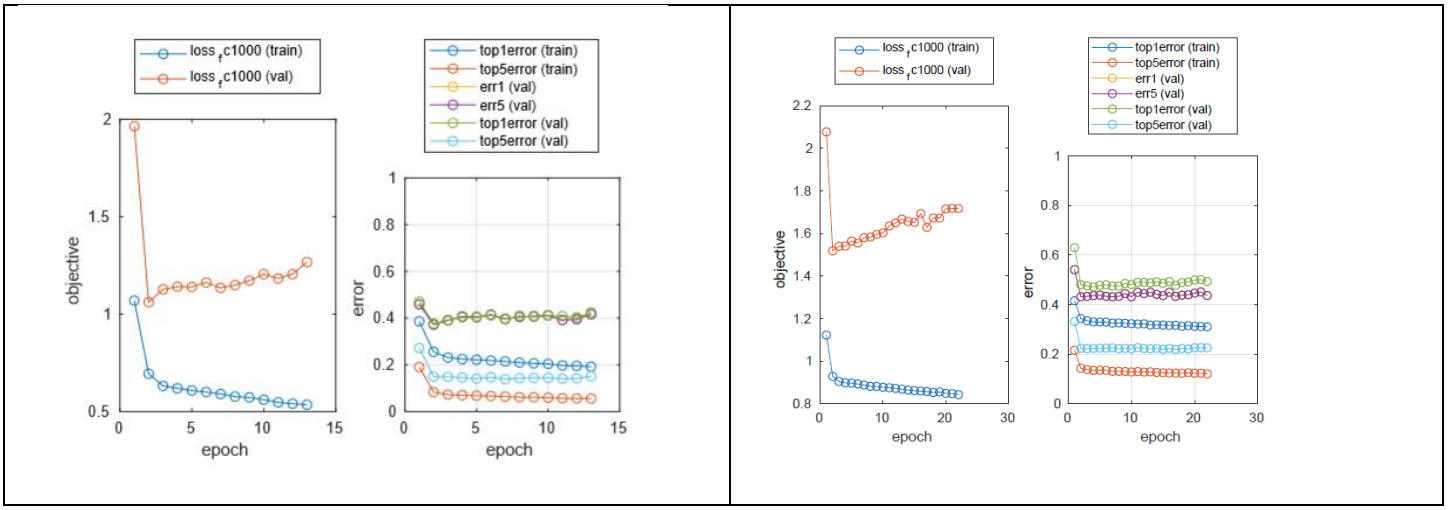
Severity		1	2	3	4	5	Average (%)
Level-1	**Sensitivity**	0.8000	0.6470	0.7173	0.7534	0.9411	**77.17%**
	Specificity	0.9869	0.9337	0.8688	0.9052	0.9735	**93.35%**
Level-2	**Sensitivity**	0.9333	0.7058	0.7826	0.9041	0.8823	**84.16%**
	Specificity	0.9803	0.9668	0.9754	0.9157	0.9602	**95.96%**

While this probability calculation appears to be a better indicator of the final classification results for each dataset, it does not apply to the trained model that relies purely on 2D spatial slices, e.g. ResNet-50. This is because, for depth-Resnet, during the training, each block has already embodied the depth information and focuses on the most discriminating patterns. On the other hand, for a 2D slice-based model, e.g. ResNet-50, all those less severe slices that scored 4 and 5 for each high severity block will be included and calculated individually. This will contribute to the final calculation of probability considerably. Hence, to compare the performance of the Resnet-50 model and depth-Resnet as developed in this study, a calculation of volume scores that are based only on block scores was created (Level-0), which applies a SVM classifier. Table 4 gives the classification results with both the depth-Resnet model enhanced in this study and the resnet-50 model trained on 2D spatial slices only. In addition.

Table 4. Comparison of scoring results applying both depth-Resnet model and ResNet-50 based on block scores.

Severity	**1**	**2**	**3**	**4**	**5**	**Average**
Depth-ResNet (Level-0)	0.93 ± 0.11	0.66 ± 0.08	0.85 ± 0.03	0.92 ± 0.03	0.85 ± 0.24	**92.70 ± 0.97%**
ResNet-50	0.60 ± 0.20	0.33 ± 0.21	0.75 ± 0.08	0.67 ± 0.01	0.66 ± 0.08	**67.15±1.69%**

The learning information for both Depth-ResNet and ResNet-50 models is exemplified in Figure 7, where the learning rate is set to 10^{-2} ; that is, decreased by an order of magnitude after the validation error saturates. The batch size is 128 slices or 16 blocks (with 8 slices each), randomly selected from all five categories.



(a) depth-ResNet	(b) ResNet-50

Figure 7. Learning information on depth-ResNet (a) applied in this study and Resnet-50 (a) .

It appears that at epoch 3 in Figure 7, the convergence takes place for both models, whereas the errors remain similar for the following epochs.

In general, the injection of depth information into the training network produces much better performance than the training purely based on 2D spatial slices, with an average classification accuracy of 92.70% in comparison with 67.15% for depth-Resnet and ResNet-50 modes respectively. Although severity 1 has the least number of datasets with 15 (10 of which were training sets and 5 testing sets), this class has been identified well with 93% accuracy for the Level-0 measure, the best among the 5 categories. This could be explained by the fact that this class has very distinctive patterns; significantly different from the others as demonstrated in Figure 1, displaying the most serious TB conditions spreading to nearly every slice of a volume. With regard to the level of severity, class 2 (Severity 2) appears to be the most challenging one to predict, not only because it has a small number of samples (17 in total), but also because its patterns bear similarities to either Severity 1 or Severity 3. This trend also occurs for the Level-1 and Level-2 calculations in Table 2, with the least accuracy of 60% and 70% realised respectively. Although class 5 (Severity 5) also shares smaller sample size (n=17), the abnormalities presented in the images as manifested in Figure 1 appear to be the least (i.e. the closest to normal), a characteristic that is much more distinguishable than in some of the other classes, which in part contributes to higher classification accuracy for both Level-1 (82%) and Level-2 (84%).

5. Discussion

This research utilises state of the art ResNet deep learning techniques to classify severity scores of tuberculosis (TB) disease from 3D CT images and has demonstrated an overall accuracy of 85.29% when taking into account severity probability and 92.70% for classification of severity scores. Due to the shortage of training datasets (170 in total), segmenting a whole volume into sub-blocks appears to be a better way forward not only to enlarge the datasets for training but also to allow the model to concentrate on those discriminative patterns between classes, since many diseased regions only occur in small blocks. However, while each block has been well trained for classification, the calculation of an overall score for a volume needs to be addressed. This study maps each score with equally spaced numbers working as the probability of high severity within the range of 0 and 1, i.e. 0.9, 0.7, 0.5, 0.3, 0.1 for scores of 1 to 5 respectively to calculate three measures, $probability_{high}$, $probability_{low}$ and $probability_{all}$ representing high, low and overall probability of a dataset with multiple sub-block scores. As a result, three approaches are developed. Level-1 predicts an overall probability based on $probability_{all}$ whereas

Level-2 delivers predictions according to $probability_{high}$ and $probability_{low}$. Based on the number of scored blocks only, Level-0 determines the overall score. In all these integrations, SVM based classifiers are applied. While the block scores remain the same, the averaged predictions for volumetric datasets of five classes are 92.70%, 75.88% and 85.29% respectively for the approaches of Level-0, Level-1 and Level-2. Although Level-0 appears to perform the best, it does not have associations with high severity probabilities. To this end, Level-2 tends to perform better.

In this research, simple linear mapping was employed. When the inverse severity score is also evaluated, i.e. using $[\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}]$ to replace [0.9, 0.7, 0.5, 0.3, 0.1], the averaged accuracy for Level-2 is 66.17% (47% for Level-1) with better performance realised for low severity TB (scored 4 and 5). Since Eqs. (14) to (16) are not independent of block scores, combining all measures (e.g. block scores, $probabilityhigh$, $probabilitylow$, $probabilityall$) provides similar results with an averaged accuracy of 91.17% for classification of severity scores. Therefore, incorporation of medical knowledge is necessary to map severity scores (1 to 5) with high severity probability, which in the future will be allied with more sophisticated or non-linear mapping formulae through deep learning techniques.

5.1 Addressing overfitting issues

Understandably, training depth-Resnet is even more prone to overfitting than training spatial ResNet. This was addressed by employing depth frame jittering. In each training iteration, 8 frames are sampled from each of the training volumes in a batch by randomly sampling the starting frame, and then randomly sampling the depth stride (between 1 and 7). Whilst during testing, a batch of samples, each comprising 8 equally spaced frames from a volume was selected, which were then propagated through the net to yield a single prediction for each volume. Instead of cropping the image corners, centre and their horizontal flips, a faster fully convolutional testing strategy as discussed in [30, 31] was applied to the original images and their horizontal flips, and the predictions from all locations were averaged. Thus inference can be performed in a *single* forward pass for the whole volume.

5.2 Comparison with the state of the art on scoring TB severity

Clinically, the assessment of the TB severity score is determined based on a number of clinical information sources, as addressed in [32], whereby, in addition to images, both clinical and laboratory data are applied, including drug resistance, presence of TB symptoms, etc.. As a result, in [32], the best prediction result is realised using combined datasets with r=0.619 and RMSE=0.791 (Root Mean Square Error). It is envisaged that both enlargement of datasets and inclusion of clinical and laboratory data could lead to more accurate severity scoring. During the imageCLEF 2018 Competition on Tuberculosis Tasks [33], a number of approaches were developed, using only CT images, including both hand-crafted and unsupervised machine learning methods. Given the small sample size of training datasets, hand-engineered models usually work better. With regard to RMSE, in this

competition, the best result (RMSE=0.7840, AUC=0.7025) was delivered applying a lesion-based TB descriptor and a random forest (RF) classifier, by incorporating age and gender information [34]. In terms of accuracy, the best result was achieved using traditional texture-based graph models [35] with AUC being 0.7708 (RMSE=0.8934). To this end, the team employed conventional approaches for the extraction of quantitative image descriptors, such as statistical moments, fractal dimension, gray-level co-occurrence matrices and their derivative features. Similarly, a texture-based 3D model was applied by another group [36], employing a range of statistical measures (mean, skewness, kurtosis, homogeneity, energy, entropy, fractal dimension). With regard to AUC, their result achieved 21th position (AUC=0.6239). However, with regard to RMSE (=0.8883), the group ranked in 3rd position. Another traditional hand-crafted method, the feature-based approach [37], appeared to also perform well for scoring the severity of TB, through the employment of image binarization and extraction of features (calcifications, lung wateriness, cavities, infection ratio, HU histograms and lung shapes), with AUC=0.6862 (8th position) and RMSE=1.1046 (25th position). In the competition, this paper appeared to be the only one applying an unsupervised approach to scoring severity. While a number of deep learning techniques, e.g. ensemble 3D CNN, transfer learning, have been developed, they are employed mainly for TB classifications. As such, Level-1 of the unsupervised approach was submitted [38] by the authors and achieved AUC=0.6534 (14th position) and RMSE=1.0921 (ranked 24) among 36 submissions.

5.3 Major contribution of the paper

This paper presents the use of an enhanced depth-Resnet deep learning network to address severity scoring for TB. In essence, depth-Resnet architecture remains a 2D network but incorporating 3rd dimensional information as illustrated in Figure 3(b). In this way, full advantage can be taken of the 2D Resnet network. As discussed in Section 5.1, random sampling takes place along the depth direction, to avoid overfitting. Although three-dimensional CNN have been employed in studying TB data in a number of research realms, such as in detection of lung nodules [39, 40], for the scoring of TB severity that is conducted in this paper, 3D CNN appears to suffer severe overfitting, which was the reason that depth-Resnet was proposed. In comparison with Resnet-50, at each layer, the 2D feature maps in the proposed depth-Resnet architecture contain not only spatial information as calculated in Resnet-50, but also depth information along the z-direction as calculated in Eq. (10), which constitutes a major contribution of the paper.

5.4 Comparison with other public TB datasets

The results presented in this paper are based on the dataset published in the ImageCLEF competition [41, 42] on TB tasks. As discussed in the above section, the classification of TB severity scores cannot be made based on CT images alone. Other complementary information (clinical and laboratory data) should be also taken into account.

However, if more accurate information can be revealed from imaging datasets, then more accurate diagnosis can be delivered. Hence, in the future, comparison with other datasets available in the public domain will be conducted, including Kaggle [13], TB annotation [43], JSRT [44], and ANODE09 [45]. Most of these datasets are utilised for detection of TB nodules [46] and should further improve the accurate detection of abnormal patterns, which will manifestly benefit severity scoring, leading to revealing underlying connections.

5.5 Comparison with a 3D CNN network

Although CT pulmonary images are in three-dimension (3D), the abnormal features to be studied at many cases are in 2D form. For instance, Figure 10 demonstrates an example of Severity 2 TB depicting middle frames (n=56) where red circles indicate abnormality. Not only a very small number of frames (7/121) contain diseased patterns along depth direction, but also those abnormal patterns occupy small spaces in a 2D frame (~1%). Hence many sub-volumes created from this dataset (with a sampling size of 128×128×30 voxels to be detailed below) contain only a few slices with abnormal features whereas most of the sub-volumes present normal patterns, making the abnormal feature volumes even smaller. Therefore, it is expected that a conventional 3D CNN network will not perform as well as a 2D CNN network. Since the developed depth-Resnet in this paper in essence is of a 2-dimensional network, it is worth well to compare with a conventional 3D CNN architecture for future enhancement.

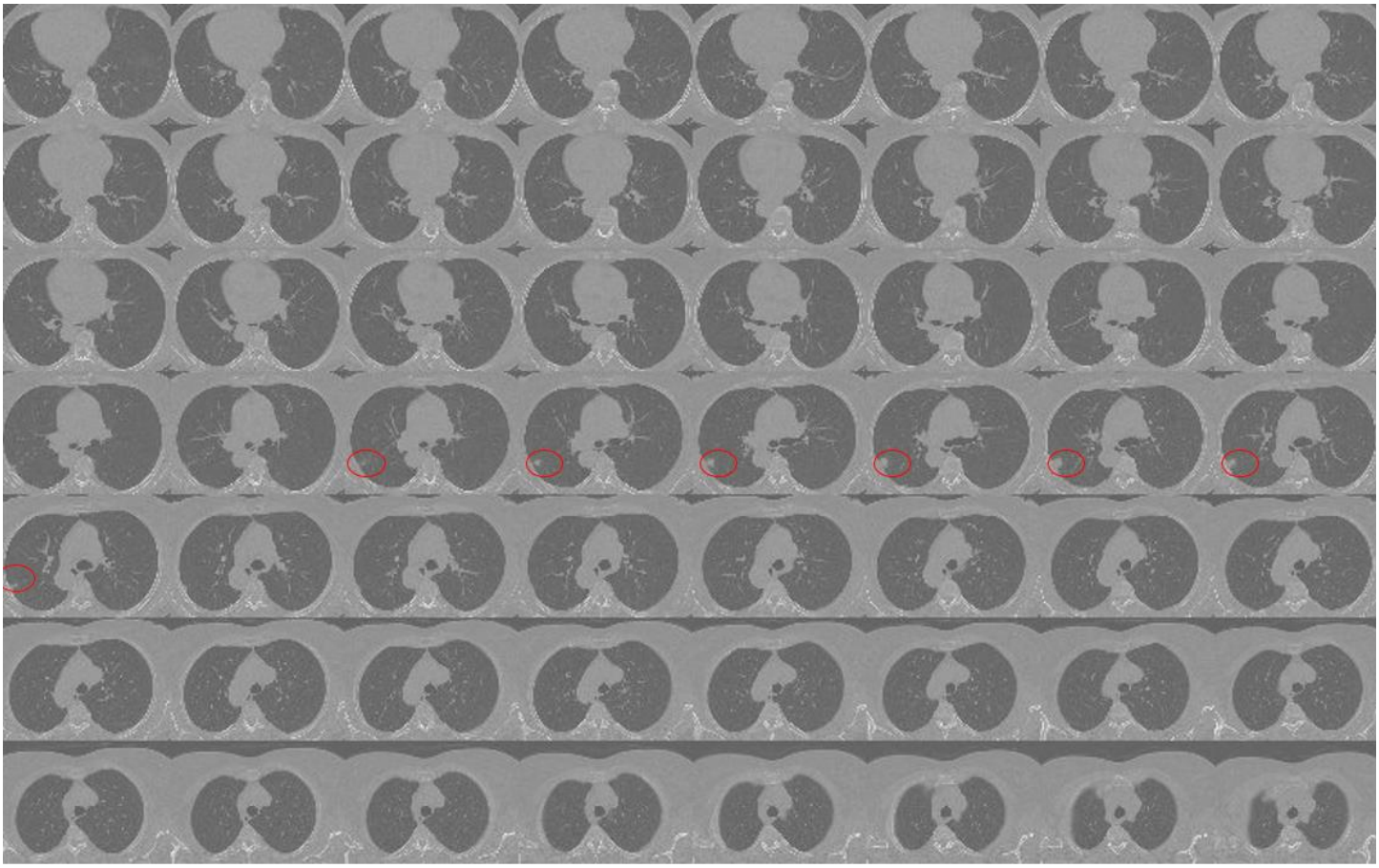


Figure 8. The abnormal pattern of an example with Severity 2 of TB with circles indicating abnormality.

In this study, the 3D CNN architecture is similar to the 3D network employed in [16] as illustrated in Figure 9. The learning rate is set to be 0.01. This 3D CNN network comprises 6 convolutional layers and one fully connected layer with detailed parameters given in Figure 9. For example, the convolution layer 1 (Conv-1 in Figure 9) has a filter size of (8, 8, 4) with 96 kernels. The stride is (2, 2, 2) with 0 padding. This layer of Conv-1 is then followed by a pooling layer with pooling size and stride being (2, 2, 1) and (2, 2, 2) respectively.

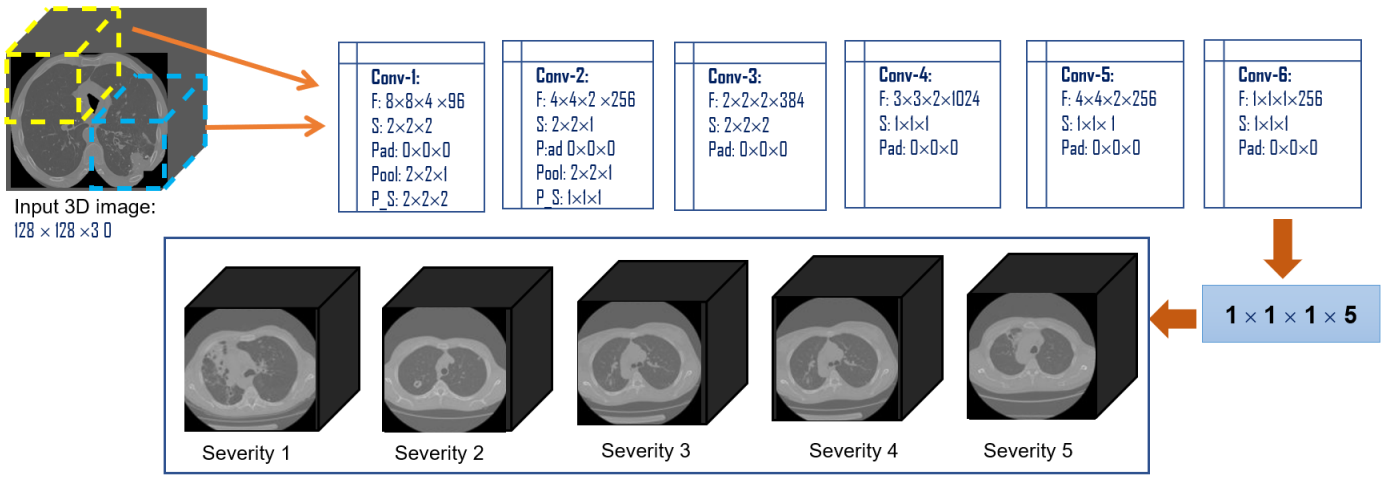


Figure 9. The architecture of the applied conventional 3D CNN network.

Each 3D volume/block was normalised to a dimension of 128×128×30 voxels generated from the blocks listed in Table 1, with a maximum of 10 slices overlapping in the depth direction. Table 5 details the number of volumes being trained and tested whereas Table 6 lists the classification results together with sensitivity and specificity information.

Table 5. Detailed information for training a 3D CNN.

	High Severity			**Low Severity**		**Total**
Severity	1	2	3	4	5	
Train (subject)	7	7	21	28	7	70
Train (volume)	400	293	715	1780	365	3553
Validation (subject)	3	3	9	12	3	30
Val (volume)	126	158	249	631	189	1353
Test (subject)	5	7	18	33	7	70
Test (volume)	314	356	406	1832	327	3235
Total	15	17	48	73	17	170
	840	807	1,370	4,243	881	8,141

Table 6. The scoring results (Level-2) for the applied 3D CNN network together with their corresponding sensitivity and specificity.

Severity	**1**	**2**	**3**	**4**	**5**	**Average**
3D CNN	0.4 ± 0.2	0.285 ± 0.143	0.176 ± 0.0	0.205 ± 0.03	0.142 ± 0.0	**24.16 ± 4.6%**
Sensitivity	0.4 ± 0.2	0.285 ± 0.143	0.176 ± 0.0	0.212 ± 0.0	0.142± 0.0	**24.16 ± 4.6%**
Specificity	0.710 ± 0.07	0.725 ± 0.0	0.711 ± 0.0	0.93 ± 0.01	1.0 ± 0.01	**81.24 ± 1.6%**

As it can been seen in Table 6, based on the level-2 calculation, the average accuracy for scoring five severity classes is 24%, which is much lower than applying depth-ResNet (85%). While the sensitivity is low (24%), the specificity appears to be high with an average of 81%, specifically for low severity of Severity 4 (93%) and Severity 5 (100%), indicating that the low-severity data are more likely to be rejected correctly. The standard deviation was calculated based on two runs.

Figure 10 depicts the learning information and shows the network appears to not converge well, with training curve (dashed line around error 0.8) tending not to change. As a result, only level-2 calculations are given in Table 6 since for level-1 calculation, there are not clear boundary lines (thresholds) between 5 classes, i.e. every volume/block scores similar in the range of [0.368, 0.610].

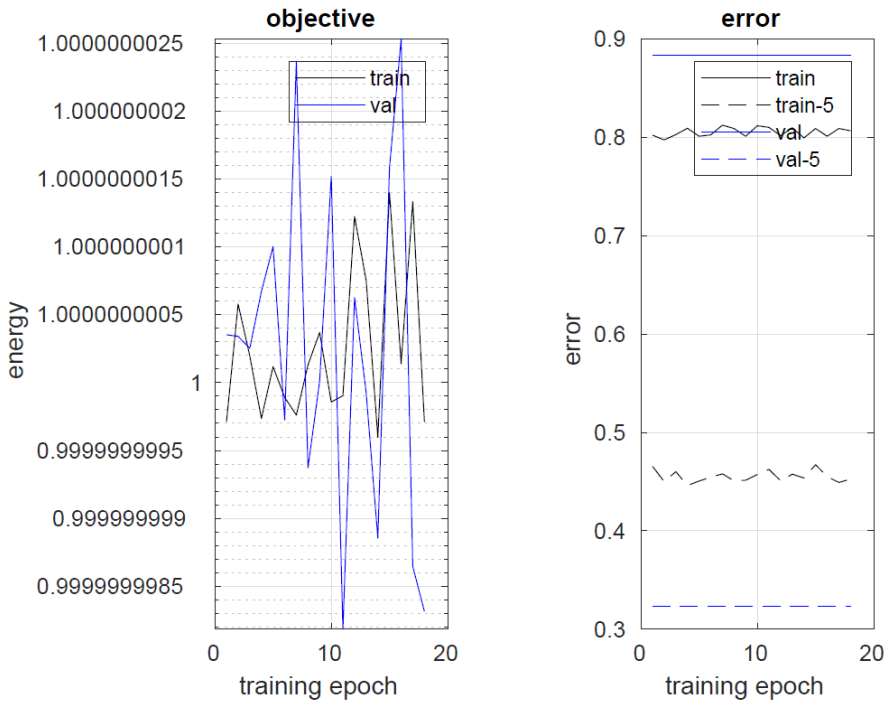


Figure 10. The learning information for a 3D CNN network.

While data sparsity (n=70 for training) contributes to this poor performance, another reason is the visual similarity of abnormal patterns across five classes. Different from the work that is carried out in [16] where concerned three classes sustain more distinguishable visual patterns (tumour, Alzheimer's disease, normal), similar patterns of TB (e.g. military, infiltrative) can take place at any level of severities. Although this 3D CNN network in Figure 9 can be further enhanced into going deeper (e.g. 50 layers), it suffers vanishing gradient problem as being addressed in Section 2.2 and illustrated in Figure 10. In [16], while the best accuracy result of 87% is achieved for classification of three classes, the architecture applied is the fuse of both 2D and 3D networks. The 3D network alone did not deliver better results than applying only 2D CNN as well in [16].

6. Conclusion

This research investigates the feasibility of applying deep learning techniques to the analysis of the level of severity of tuberculosis disease from 3D CT images. To this end, the developed depth-Resnet deep learning model was trained based on segmented blocks from each data set. The final severity score for each volume was therefore built on the integration of segment scores. As such, three approaches were presented to assemble the integration. One was constructed based on the block scores using a SVM classifier (Level-0). Another approach was to convert block severity scores into a probability map to provide a range of probabilities for each severity score. From this, three measures were formulated, which are based on the high severity scores (1 to 3), low severity scores (4 and 5) and all scores (1 to 5). As a result, two more classifiers were developed, applying either the measure of all scores (Level-1) or two separate measures (Level-2). It appears that Level-2 approach performs the best in terms

of both severity scores and probability (85.29%). However, if only the severity score is considered, which is the information provided with the training datasets, Level-0 performs the best (92.70%). Due to the lengthy training (~4 days) and testing (~2 days) times, the results in this paper are based on three training runs. In the future, the popular method of one against all will be investigated to obtain more accurate predictions of the level of severity. In addition, better mappings from severity scores (1 to 5) to high severity probabilities (1 to 0) will be sought, to unearth the information that is hidden within the severity scores, through the training of hyperparameters directly from the datasets. While Resnet-50 was the chosen model as a baseline to be applied in this study, comparison with the other pre-trained deep learning networks (e.g. Inception-v4) and hand-crafted models (e.g. SIFT, LBP) will be investigated in the future.

Due to the sparsity of the datasets and the characteristics of similar abnormal patterns of TB within the five severity levels, conventional 3D CNN architectures appear to work less effective, which however, can be improved by training segment volumes that only contain abnormal patterns. Hence another future work includes collaborations with clinicians to incorporate expert knowledge by training only those diseased regions associated with the severity of TB.

Scoring TB severity constitutes one of the biggest challenges in medical decision-making. This work is anticipated to make a significant contribution to this field and promote the application of machine learning techniques within the medical domain.

Acknowledgement

The authors would like to thanks imageCLEF Tuberculosis organisers to providing these valuable sets of CT data.

References:

[1]. Szegedy C, Ioffe S, Vanhoucke V., Inception-v4, inception-resnet and the impact of residual connections on learning . 1026; ,*arXiv preprint* arXiv:1602.07261.

[2]. WHO, Tuberculosis, Fact Sheet, March 2017. http://www.who.int/mediacentre/factsheets/fs104/en/. Retrieved in June 2017.

[3]. Jeong YJ, Lee KS, Pulmonary tuberculosis: up-to-date imaging and management. *AJR Am. J. Roentgenol*. 2008*;* 191*:* 834–44.

[4]. Lange C, Mori T, Advances in the diagnosis of tuberculosis. *Respirology*. 2010; 15: 220–240.

[5]. Lee KS, Im JG, CT in adults with tuberculosis of the chest: characteristic findings and role in management. *AJR Am. J. Roentgenol*. 1995; 164: 1361–7.

[6]. McGuinness G, Naidich DP, Jagirdar J, Leitman B, McCauley DI, High resolution CT findings in miliary lung disease. J. Comput. Assist. Tomogr. 1992; 16: 384–90.

[7]. Krysl J, Korzeniewska-Kosela M, Muller NL *et al*, Radiologic features of pulmonary tuberculosis: an assessment of 188 cases. *Can. Assoc. Radiol. J.* 1994; 45: 101–7.

[8]. Fukushima K, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.* 1980; 36: 193–202.

[9]. LeCun Y, Bottou L, Bengio Y, and Haffner P. Gradient-based learning applied to document recognition, *Proceedings of the IEEE 1998*; 86(11): 2278–2324.

[10]. LeCun *Y*, Bengio Y, Hinton G, Deep Learning, *Nature. 2015; 521: 436-444.*

[11]. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems 2012.* NIPS 2012.

[12]. ImageNet, http://www.image-net.org/.

[13]. Kaggle Competition, https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15801.

[14]. Pereira S, Pinto A, Alves V, and Silva C. Brain tumour segmentation using convolutional neural networks in MRI images, *IEEE transactions on medical imaging 2016*; 35(5):1240-1251.

[15]. Gao X, Li W, Loomes M, Wang L, A fused deep learning architecture for viewpoint classification of echocardiography, *Information Fusion 2017; 36:103-113.*

[16]. Gao X, Hui R, Tian Z, Classification of CT images based on deep learning networks, *Computer Methods and Programs in Biomedicine 2017;* 138:49-56.

[17]. Gao, **X.,** Quan, Y., Application of Deep Learning Neural Network for Classification of TB lung CT Images based on Patches, *CLEF 2017 CEUR Working Notes,* Vol-1866, 2017.

[18]. Gao, X**.,** Quan, Y., Prediction of multi-drug resistant TB from CT pulmonary Images based on deep learning techniques, *Molecular Pharmaceutics 2018; in press.*

[19]. 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 Conference on Computer Vision and Pattern Recognition*, 2015, pp 1-9.

[20]. He K, Zhang X, Ren S, Sun J, Identity Mappings in Deep Residual Networks, *European Conference on Computer Vision (ECCV)*, 2016.

[21]. He K, Zhang X, Ren S, Sun J, Deep Residual Learning for Image Recognition, IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22]. Ioffe S, Szegedy C, Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In Proceedings of The 32nd International Conference on Machine Learning* 2015; pp 448–456.

[23]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, Rethinking the inception architecture for computer vision. 2015; arXiv preprint arXiv:1512.00567.

[24]. Feichtenhofer C, Pinz A, Wildes R, Temporal Residual Networks for Dynamic Scene Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[25]. Feichtenhofer C, Pinz A, Wildes R, Spatiotemproal residual networks for video action recognition**,** in Proc. NIPS 2016

[26]. MatConvNet: http://www.vlfeat.org/matconvnet/. Retrieved in May, 2018.

[27]. *Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20 (3): 273–297.*

[28]. Cappellato L., Ferro N., Nie J, Soulier L, Eds., CLEF 2018 Working Notes, Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, CEUR-WS, Eds., 2018.

[29]. ImageCLEFtuberculosis, http://www.imageclef.org/2018/tuberculosis. Retrieved in May, 2018.

[30]. Simonyan K, Zisserman A., Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*, 2014.

[31]. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *In Proc. ICLR*, 2014.

[32]. 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. In: Computer Assisted Radiology and Surgury - 32nd International Congress and Exhibition (CARS-2018). Volume 13., Springer, Heidelberg (2018) 143–144.

[33]. 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, CLEF working notes, CEUR, 2018., CEUR-WS.org, September 10-14, Avignon, France (2018).

[34]. Liauchuk, V., Tarasau, A., Snezhko, E., Kovalev, V., Gabrielian, A., Rosenthal, A.: ImageCLEF 2018: Lesion-based TB-descriptor for CT image analysis. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018).

[35]. Dicente Cid, Y., M¨uller, H.: Texture-based graph model of the lungs for drug resistance detection, tuberculosis type classification, and severity scoring: Participation in ImageCLEF 2018 tuberculosis task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018).

[36]. Ahmed, M.S., Obaidullah, S.M., Jayatilake, M., Gon ̧calves, T., Rato, L.: Texture analysis from 3D model and individual slice extraction for tuberculosis MDR detection, type classification and severity scoring. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018.

[37]. Bogomasov, K., Himmelspach, L., Klassen, G., Tatusch, M., Conrad, S.: Feature-based approach for severity scoring of lung tuberculosis from CT images. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.orghttp://ceur-ws.org (September 10-14, 2018).

[38]. Gao, X., James-Reynolds, C., Currie, E.: Scoring TB severity with an enhanced deep residual learning depth-resnet. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org http://ceur-ws.org (September 10-14 2018).

[39]. Huang X., Shan J., Vaifya V., Lung nodule in CT using 3D convolutional neural networks, IEEE 14th Internal Symposium on Biomedical Imaging (ISBI 2017), 2017.

[40]. Liao F., Liang M., Li Z., Hu X., Song S., Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network, ArXiv: 1711.08324. https://arxiv.org/abs/1711.08324.

[41]. Dicente Cid, Y., , Liauchuk, V., Kovalev, V., , M¨uller, H.: Overview of Image-CLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org http://ceur-ws.org (September 10-14 2018).

[42]. Ionescu, B., M¨uller, 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), Avignon, France, LNCS Lecture Notes in Computer Science, Springer (September 10-14 2018).

[43]. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017.

[44]. JSRT: http://db.jsrt.or.jp/eng.php. Retrieved in August 2018.

[45]. B. van Ginneken, S.G. Armato, B. de Hoop, S. van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A.M.R. Schilham, A. Retico, M.E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Belloti, F.D. Carlo, R. Megna, S. Tangaro, L. Bolanos, P. Cerello, S.C. Cheran, E.L. Torres and M. Prokop. "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study", *Medical Image Analysis* 2010;14:707-722.

[46]. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, and Doi K.: Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR 174; 71-74, 2000.