Deep Learning Models for Tuberculosis Detection from Chest X-ray Images

Quang H. Nguyen[1], Binh P. Nguyen[2], Son D. Dao[1], Balagopal Unnikrishnan[3], Rajan Dhingra[3], Savitha Rani Ravichandran[3], Sravani Satpathy[3], Palaparthi Nirmal Raja[3] and Matthew C. H. Chua[3]
[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[2]School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand
[3]Insitute of System Science, National University of Singapore, Singapore, Singapore
Email: [1]quangnh@soict.hust.edu.vn, [2]b.nguyen@vuw.ac.nz, [3]mattchua@nus.edu.sg

***Abstract*—This paper explores the usefulness of transfer learning on medical imaging for tuberculosis detection. We show an improved method for transfer learning over the regular method of using ImageNet weights. We also discover that the low-level features from ImageNet weights are not useful for imaging tasks for modalities like X-rays and also propose a new method for obtaining low level features by training the models in a multiclass multilabel scenario. This results in an improved performance in the classification of tuberculosis as opposed to training from a randomly initialized settings. In other words, we have proposed a better way for training in a data constrained setting such as the healthcare sector.**

***Index Terms*—Tuberculosis, detection, classification, X-rays, transfer learning, deep learning, medical imaging**

I. Introduction

Tuberculosis is a bacterial disease primarily affecting the lungs. It is a curable and preventable disease affecting developing nations and is one among the top ten causes of death worldwide. The early identification and treatment of tuberculosis is of great value both in terms of reducing cost of treatment and improving health outcomes [1]. The magnitude of the problem can be inferred from the fact that the World Health Organization has a separate EndTB strategy which aims at reducing deaths caused by tuberculosis and the incidence of tuberculosis by 95% and 90%, respectively, by 2035 [2].

In the detection of tuberculosis, there are two pathways that are relevant. One is the patient-initiated pathway where improved awareness of symptoms among people can help early detection and the second pathway being the screening pathway where low cost screenings are required to be systematically done in patient populations of high risk [3]. In the screening process, medical imaging plays a significant role. Chest X-rays are useful in non-invasive diagnosis and screening tools [4].

Tuberculosis is important to be tackled at a global level and is also important for meeting the United Nations sustainable development goals and is of great importance to developing nations as they affect mainly the working population. The development of medical imaging techniques and algorithms for tuberculosis detection is hence of great importance – as the reduction in incidence of tuberculosis and its elimination in turn reduces global poverty and improves healthcare outcomes for people of developing nations.

For our work, we seek to look at tuberculosis detection for the large-scale screening from chest X-rays. The systems developed would be able to help in multiple ways. Primarily, by detecting the percentage chance of tuberculosis given a chest X-Ray, it would be possible to determine if more in depth tests are required for the confirmation of the disease. The visualisation makes the models more interpretable and reduces the clinician's overall workload and thereby is useful as a good augmentation strategy in developing nations where healthcare professionals are few in number and stretched on resources.

II. Dataset and Related Work

For the purpose of our experiments, we rely mainly on publicly available datasets. The following are the datasets being considered Montgomery and Shenzhen datasets [5] and NIH-14 dataset [6] hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.

Shenzhen dataset [7] consists of chest X-rays taken from Shenzhen No. 3 Peoples Hospital which are 4020×4892 pixels in dimension. Montgomery dataset consists of frontal chest X-rays taken from the Department of Health and Human Services in partnership with Montgomery County in the United States. These were taken for the purpose of screening of tuberculosis and is similar in dimension to Shenzhen dataset. For the purpose of the transfer learning experiments we have taken NIH-14 dataset [6] which is a collection of chest X-rays from clinical PACS databases coming under the National Institutes of Health Clinical Centre. It consists of 112,120 chest X-ray images with labels consisting of 14 common chest pathologies. These do not include tuberculosis as a label and is useful for learning the lower level features while transfer learning is done.

Jaeger *et al.* [5] extracted several features from lung segmented CXRs and employed various classification methods to benchmark the features. Rajpurkar *et al.* [8] has previously shown that NIH-14 dataset can be tackled as a classification problem and has provided useful ways of modelling the problem to be solved by deep convolutional networks. Hwang *et al.* [9] have previously worked on Shenzhen dataset. The usage of class activation maps (CAMs) are shown to be useful

for visualisation of convolutional networks by [10]. For the purpose of training, we have augmented the datasets in a manner that is consistent to the ways in which an X-Ray maybe distorted. Rotations, horizontal flipping and perspective transforms are used as augmentation strategies which mirror the real-world scenarios of image flipping, image skew and flipping which occur while scanning of X-rays. A simple rescaling operation is done before feeding to the neural networks to bring the images in the numerical range zero to one. Even though contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE) was experimented with, as it did not provide any improvement in obtained results.

III. Experiments

For the detection of tuberculosis, the task is considered to be a binary classification task with the final layer giving the probability that the given X-rays image has tuberculosis. The given set of experiments must deal with answering the following – the effectiveness of using ImageNet weights for transfer learning and methods to improve that performance using pre-trained models. Since the final aim is to develop a system that can be deployable and can be generalized in a much broader sense, in addition to the above, we need to also look at the ideal architectures that can be used, the elements within architectures that are detrimental to the performance for the given task, the selection of loss function such that it can be used for a variety of diseases and finally make the models interpretable for clinical use. The pipeline of the experiments is showed in Fig 1.

For comparison of the backbones to be used in the models, we first train models in the following manner – keeping the set of images in Shenzhen dataset for training and validation purposes and testing the models on Montgomery dataset. In these experiments, keeping the all other parameters such as learning rate, early stopping and the train-validation splits to be constant, we vary the backbone of the architecture from among a set of proven architectures: ResNet-50, VGG16, VGG19, DenseNet-121 and Inception ResNet.

Here the models are initialized with ImageNet weights and the training is done for 100 epochs on each model. From this we select the best architecture to be used in the next step of the process. In the next step as well, we keep the train-validation splits same and keep Montgomery dataset as the test set. In addition, we also require to understand how the connection from the convolution blocks to the dense layers affect the performance. Hence on the best model chosen, we run a set of experiments by connecting the convolution block which gets the image features to the prediction block which is a fully connected layer with a single node. We connect these, first with a set of two fully connected dense layers and then directly by taking the average of the final convolution layers in the following fashion:

$$Res_{x,y} = \sum_{k=1}^{d} \frac{f_{x,y,z}}{d} \forall x \in X, y \in Y, \tag{1}$$

where
d = Number of feature maps,
$f_{x,y,z}$ = Value at position x, y, z of a feature map,
X = [1, height of function],
Y = [1, width of function].
On comparing these, we see which elements are prone to over-fitting and which is best for generalization in our scenario.

Given, the best architecture setting, we see how well the images perform with the ImageNet weights. We also see till whether the features learned on the ImageNet dataset is useful by unfreezing the layers and looking for an increase or decrease in performance. If the model's performance increases as more and more layers are unfrozen, we can conclude that the learned weights on ImageNet are not useful and that the model has to learn a new set of features to better cope with the task at hand. Exploring further to see if another set of pre-trained weights can be used, we look at NIH-14 dataset as an ideal dataset to do the pre-training. The dataset has a variety of diseases (among which tuberculosis is not present) and is from the same modality as that of the data under consideration for tuberculosis. However, the training cannot be done in the same manner as it is no longer a binary classification problem. Hence, we change the last layer to have 14 nodes to indicate each of the diseases and train the images with the same train-test splits as that of [11]. The loss function is now varied to accommodate for multi-class and multi-label classification. This proves to be useful finding the generic loss function that we were looking for – so that the system can generalize to other cases as well. The loss per disease function is now set as:

$$LPD_d = -y_d \log P(Z) - (1 - y_d) \log P(\bar{Z}) \tag{2}$$

where y_d is the target disease, Z is the event that presence the disease and $\bar{Z}$ is the event that absence the disease. The total loss function is as follows:

$$TotalLoss = \sum_{i \in D} LPD_i \tag{3}$$

where D is the set of Disease. With the metrics remaining the same as before for the tuberculosis classification task. Area under the Receiver Operator Curve (AUROC or AUC) is used as the evaluation metric. Here the only difference is that, in a binary setting this is a single value, in a multi-class setting, we take per-class AUCs.

Once the training is done on NIH-14 dataset, we take the convolution layer weights and revert to our old setting of binary classification setting, now using these weights with the architecture obtained from the initial experiments. Now we compare the effectiveness of pre-training on ImageNet vs. NIH datasets. Under identical training settings, we look at the AUC scores while we unfreeze more layers for training. If our hypothesis, that NIH weights are better and they help learn the lower layer features better, we should see two trends – the first being that the model with NIH weights should perform better than ImageNet weights and the second being that if NIH weights are useful, we should see a dip in performance when all layers are unfrozen.

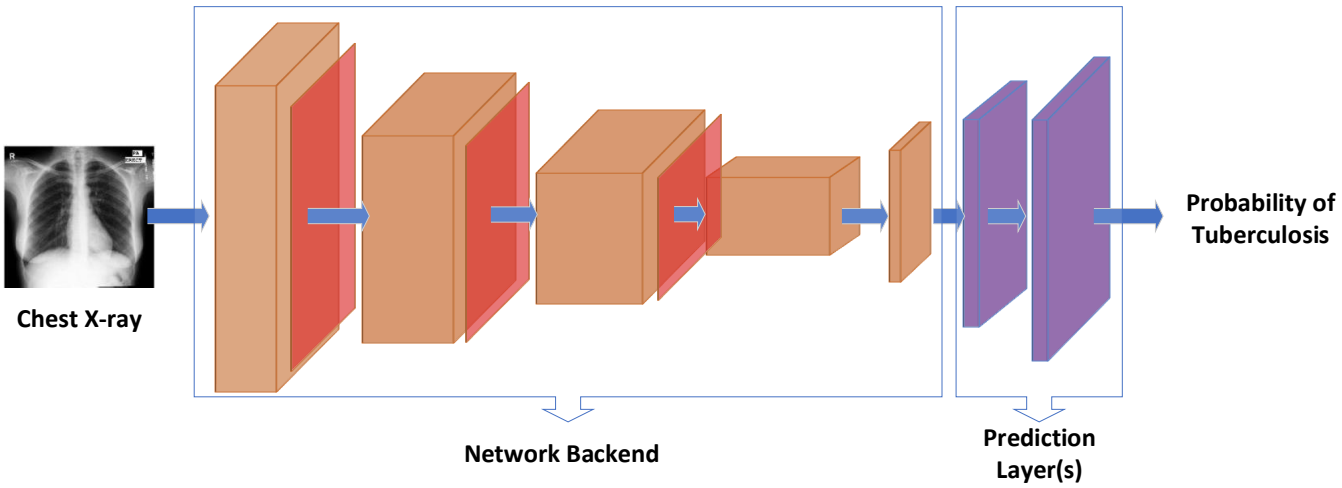

Fig. 1. Overall architecture of the propose method. A chest X-ray image is the input of a Network Backend which can be a set of existing architectures such as ResNet-50, VGG16, VGG19, DenseNet-121 and Inception Resnet. It is followed by a Prediction Layer(s) which is a combination of Average Pooling layer and Fully Connected layer(s). The Probability of Tuberculosis is the output of this network.

TABLE I
AUCs ON VARIOUS ARCHITECTURES

Model	Initialized Weights	Shenzhen AUC	Montgomery AUC
Resnet-50	ImageNet	0.99	0.70
VGG-16	ImageNet	0.50	0.50
VGG-19	ImageNet	0.50	0.50
Inception ResNet V2	ImageNet	0.99	0.80
Dense Net	ImageNet	0.99	0.80

TABLE II
AUCs ON TUNING DENSE NET MODEL

Model	Initialized Weights	Shenzhen AUC	Montgomery AUC
Freeze all	ImageNet	0.65	0.54
Unfreeze 5	ImageNet	0.65	0.70
Freeze all + 2 FCN 10	ImageNet	0.70	0.69
Unfreeze 10 + 2 FCN	ImageNet	0.80	0.72
Unfreeze all + 2 FCN	ImageNet	0.99	0.74
Global Avarage Pooling	ImageNet	0.99	0.82

TABLE III
AUCs ON TRANSFER LEARNING MODEL WEIGHTS

Model	Initialized Weights	Shenzhen AUC	Montgomery AUC
Freeze all	NIH-14	0.85	0.82
Unfreeze 5	NIH-14	0.85	0.81
Unfreeze 10	NIH-14	0.89	0.84
Unfreeze 20	NIH-14	0.89	0.83
Unfreeze all	NIH-14	0.99	0.79

The results are presented in the next section. For interpretability, we have also presented the class activation masks which have been found to be useful to visualize which portion of the image has contributed to the classification result [10], [12]. This gives the clinicians an added functionality that they can now look at which areas of the X-rays to concentrate more on and also gives an intuitive idea of how the model has interpreted the X-rays.

IV. Results and Discussions

We conducted various test involving Shenzhen and Montgomery datasets and finalized the structure which is resulting into highest Test Set Accuracy.

A. Backbone Finalization

Firstly, all the backbones were tried on the Model training (Table I). Inception ResNetV2 and Dense Net performed equally well on the Test Set (Montgomery). We can notice that performance is better by models with skip connections rather than sequential connections. VGG models do not have skip connections and perform only as good as random guessing as indicated by the 0.5 AUC scores. Also, we can see that ResNet-50 and DenseNet-121 give the best performances. DenseNet is chosen as the best among these as it achieves similar performance with three times lesser parameters than ResNet [10].

B. Tuning of DenseNet Models

Secondly, the architecture of DenseNet Model is finalized, the layers of the models if finalized as per Table II.

We see that the performance is better while performing the averaging as opposed to fully connected layers. This can be explained by [13] as the fully connected layers over-fitting to the training set and hence giving poor generalization. This is a known issue of fully connected layers. Hence, we stick to the second method.

C. Inclusion of Transfer Learning Model Weights

Thirdly, the Transfer Learning Model weights are used, these weights are then tested with freezing different set of layers and results is as per Table III.

On unfreezing the layers, we see that though the performance initially increases, it then plateaus and then decreases

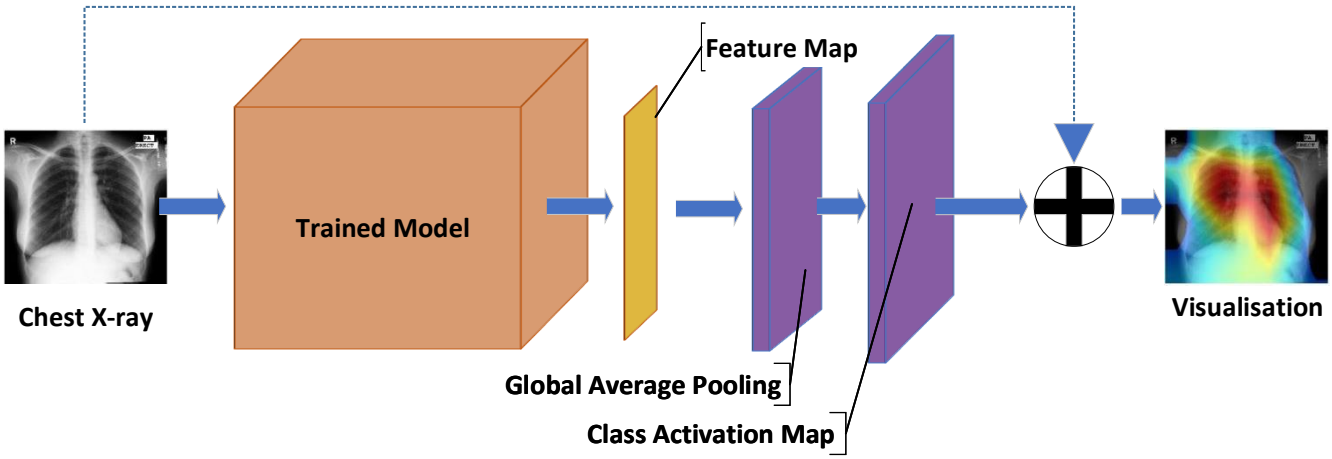

Fig. 2. Visualisation schema. The feature map (yellow block) is fed to a Global Avarage Pooling Layer and a Class Activation Map. The output of the Class Activation Map is interpolated with the input X-ray chest to make the visualisation.

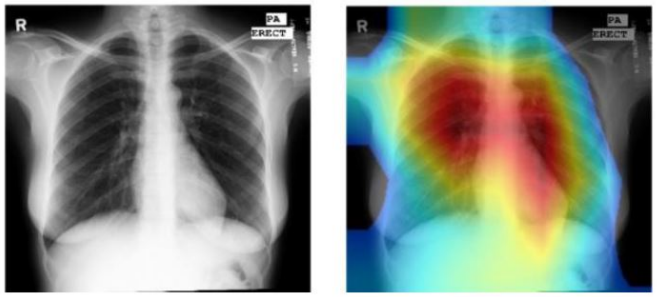

Fig. 3. Example of CAMs on the detection of normal X-rays; it focuses on the clear area in the lungs.

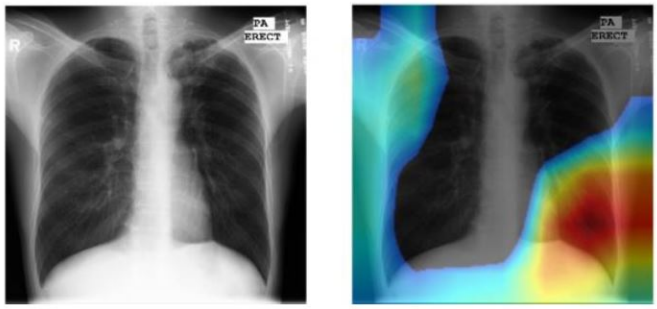

Fig. 4. Example of CAMs on the detection of abnormal X-rays; it focuses on the infected area of the lungs.

when all weights are unfrozen. This proves our initial hypothesis that these weights are useful and pre-training on the NIH dataset has helped learn useful lower level features. Unfreeze layer up to 10 resulted into best results and it was finalized for the model.

V. Model Interpretation

In order to obtain insights from the model, heatmaps were generated using class activations mappings (CAMs) to visualise the regions of the image that has the greatest resemblance of the disease [9]. To generate the CAMs (Figures 3 and 4), an image was input into the fully trained model to extract the feature maps which are then produced by the final convolutional layer. Global Average Pooling in the final feature maps helps us to focus on localization of relevant features and is expressed as class activation map. The pipeline is showed in Fig 2.

VI. Deployment and Extensibility

The deep learning model was deployed as a python web-app in Fig 5. The pipeline consisted of a python web server. The Bottle framework was chosen for this purpose as it has not external dependencies and has a one file deploy-ability which makes it ideal to package. All code has been written as a combination of TensorFlow and Keras libraries. OpenCV has been used for visualisations and IO tasks while numerical operations and optimizations have been done using numpy. In the interest of teams who might be taking the work forwards, we find the following things to be useful while the model is deployed. One single instance of a model serving results in batches is most efficient as it can use the GPU more efficiently. In the case of one at a time prediction, take care not to reload the models as it just hogs up the memory. The loading of large models to system memory is what takes the most time in these pipelines. Hence as far as possible, avoid duplicate loads or switching between models.

In certain datasets, we see that the X-Rays contain labels on its side which is indicative of the disease of the patient. This is seen when pre-diagnosed X-Rays are used for data. The model may sometimes consider them as features for prediction and such models will not generalize to the real-world cases. Hence, in this scenario a useful pre-processing technique is to use a trained U-Net to semantically segment out just the area of focus, in this case the lungs and use that for training purpose. This is useful to removing markers like labels and

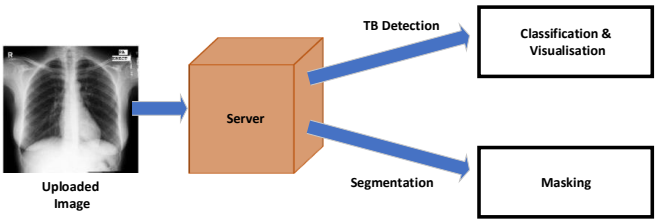

Fig. 5. Implementation of visualisation

position of certain elements which might cause the model to learn wrong features for prediction. In such cases, it would be best to visualize the model using the modules provided so see where the model focuses on most to get the predictions.

VII. Conclusion and Future Work

The experiments help understand the nature of transfer learning strategies to be used for medical images. In the process, we also develop an application for screening of patients for tuberculosis and have presented a demo consistent with problem statement presented and with appropriate visualisations. We have also suggested the ideal architectures that can be used and the elements that should be avoided to get better generalization. We show that the ImageNet weights are insufficient and the usage of appropriate data for pre-training is important and makes the entire process more efficient.

There can be multiple areas in which this work may be taken forward. One obvious question to be answered is how well the system would perform against human counterparts. A human benchmarking against several of these datasets is required to see both the individual performance and the agreement between various clinicians. From inputs of healthcare professionals, we understand that tuberculosis detection is not just limited to examination of X-rays, but can also take as input various things like patient history, lab reports and tests etc. which help improve the final prediction. An interesting area of research would be to combine these sources together and come up with an interpretable model which would make use of clinical inputs as well as the image data. Another way that may improve the performance of the system is applying ensemble learning with different classification algorithms, such as, the support vector machines, random forests, enhanced k-nearest neighbours [14], and kernel dictionary learning [15], on the features at the last fully-connected layer.

Solving the spread and incidence of tuberculosis is one which can give great rewards to developing nations. Though the problem has not been that well studied due to lack of financial motivations for private players, the release of new public datasets has helped in recent years to put this as a prominent research problem.

Acknowledgment

Q. H. Nguyen and B. P. Nguyen gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

References

[1] World Health Organisation. Global tuberculosis report 2018.
[2] World Health Organisation. Tuberculosis detection and diagnosis.
[3] World Health Organisation. Chest radiography in tuberculosis detection.
[4] World Health Organisation. Global strategy and targets for tuberculosis prevention, care and control after 2015.
[5] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3462–3471. IEEE, 2017.
[6] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona M Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer K Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014.
[7] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475, 2014.
[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225*, 2017.
[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 2921–2929, 2016.
[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 2261–2269. IEEE, 2017.
[11] Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based on deep convolutional neural networks. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97852W. International Society for Optics and Photonics, 2016.
[12] Chang Liu, Yu Cao, Marlon Alcantara, Benyuan Liu, Maria Brunette, Jesus Peinado, and Walter Curioso. TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP 2017)*, pages 2314–2318. IEEE, 2017.
[13] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv:1312.4400*, 2013.
[14] Binh P. Nguyen, Wei-Liang Tay, and Chee-Kong Chui. Robust biometric recognition from palm depth images for gloved hands. *IEEE Transactions on Human-Machine Systems*, 45(6):799–804, Dec 2015.
[15] Xuan Chen, Binh P. Nguyen, Chee-Kong Chui, and Sim-Heng Ong. Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2016)*, pages 2547–2552. IEEE, Oct 2016.