Expert Systems With Applications 184 (2021) 115519

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning

Linh T. Duong [a], Nhi H. Le [a], Toan B. Tran [a], Vuong M. Ngo [b], Phuong T. Nguyen [c,*]

[a] Institute of Research and Development, Duy Tan University, Viet Nam
[b] Ho Chi Minh City Open University, Viet Nam
[c] Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, Italy

ARTICLE INFO

Keywords:
Deep learning
EfficientNet
Tuberculosis detection
Transfer learning
Transformer

ABSTRACT

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* is a contagious disease which is among the top deadly diseases in the world. Research in Medical Imaging has been done to provide doctors with techniques and tools to early detect, monitor and diagnose the disease using Artificial Intelligence. Recently, many attempts have been made to automatically recognize TB from chest X-ray (CXR) images. Still, while the obtained performance is encouraging, according to our investigation, many of the existing approaches have been evaluated on small and undiverse datasets. We suppose that such a good performance might not hold for heterogeneous data sources, which originate from real world scenarios. Our present work aims to fill the gap and improve the prediction performance on larger datasets. In particular, we present a practical solution for the detection of tuberculosis from CXR images, making use of cutting-edge Machine Learning and Computer Vision algorithms. We conceptualize a framework by adopting three recent deep neural networks as the main classification engines, namely modified EfficientNet, modified original Vision Transformer, and modified Hybrid EfficientNet with Vision Transformer. Moreover, we also empower the learning process with various augmentation techniques. We evaluated the proposed approach using a large dataset which has been curated by merging various public datasets. The resulting dataset has been split into training, validation, and testing sets which account for 80%, 10%, and 10% of the original dataset, respectively. To further study our proposed approach, we compared it with two state-of-the-art systems. The obtained results are encouraging: the maximum accuracy of 97.72% with AUC of 100% is achieved with ViT_Base_EfficientNet_B1_224. The experimental results demonstrate that our conceived tool outperforms the considered baselines with respect to different quality metrics.

1. Introduction

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis (M.tb)*, and about 25% of the world population are supposed to be infected with *M.tb*. Every year, ten millions of people get caught by the virus, and 1.45 millions of people lost their lives because of TB. According to the World Health Organization (WHO), the epidemiological map of tuberculosis is mainly distributed in South-East Asia with 44%, followed by Africa with 24%, Western Pacific with 18%, Eastern Mediterranean with 8%, America and Europe with 3% for each. Among seven millions patients reported in 2018, 1.05 millions of them corresponding to 15%, are asymptomatic for TB. This means for these cases, it is difficult to diagnose with bacterial methods. On the other hand, among 5.9 millions patients with pulmonary TB, maximum 70% cases have been bacteriologically confirmed in recent years. As a result, the remaining cases were clinically diagnosed based on symptoms, abnormalities on CXR and medical records. Given the circumstances, a quick and accurate diagnosis would be of highly importance in the treatment and control of the disease.

Currently, the golden standard in case determination tests is still the technique for isolating bacteria. Although it has a very high specificity, the technique suffers from a relatively low sensitivity. Moreover, such a test is a long-lasting process, which may take at least three weeks to give the final results. Besides, immunological and molecular biology testing techniques have their own advantages and disadvantages. The technique of testing the molecular markers of TB in the patient samples by

* Corresponding author.
E-mail addresses: duongtuanlinh@duytan.edu.vn (L.T. Duong), lehoangnhi@duytan.edu.vn (N.H. Le), tranbaotoan@duytan.edu.vn (T.B. Tran), vuong.nm@ou.edu.vn (V.M. Ngo), phuong.nguyen@univaq.it (P.T. Nguyen).

https://doi.org/10.1016/j.eswa.2021.115519
Received 30 January 2021; Received in revised form 10 June 2021; Accepted 29 June 2021
Available online 9 July 2021

Polymerase Chain Reaction (PCR) basically overcomes the limitations of the two aforementioned methods. Unfortunately, not all medical institutions are capable of performing the PCR techniques due to the high cost.

Vietnam is a developing country, and among different issues, it still suffers from a high incidence rate and mortality caused by TB (Nguyen et al., 2020). A large number of the country's population live in remote and mountainous areas, where there is a lack of healthcare specialists and equipment. Usually, people in these regions need to travel a long distance to big cities to get checked when they have the symptom. Though the country has strengthened its system of TB diagnosis by means of various administrative efforts, fighting against the disease is still a daunting task for the authority. According to a report of the World Health Organization in 2018, Vietnam is one of the worst affected countries by TB with a high incidence rate and mortality. Fighting against TB is among the most daunting tasks for the Vietnamese healthcare authority. Currently, bacterial isolation is still the golden standard for diagnosis. Under the circumstances, there is an urgent need for pre-screening facilities to conduct early diagnosis. Among others, monitoring TB with the support of Medical Imaging techniques is beneficial to the diagnosis and prognosis of the disease.

Recently, Artificial Intelligence (AI) has gained traction in various aspects of everyday life (Alizadeh, Allen, & Mistree, 2020; Iovino, Nguyen, Salle, Gallo, & Flammini, 2021), and it has been in the forefront of methodologies applied to improve products and services (Alizadeh et al., 2019; Jain, Mishra, Shukla, & Tiwari, 2019; Zeng et al., 2021). The application of AI and Deep Learning is on the rise in healthcare (Jia et al., 2020; Soltanisehat, Alizadeh, Hao, & Choo, 2020), and various techniques have been deployed to assist doctors in their daily tasks (Sutoko et al., 1021). Several existing studies (Jiang et al., 2017; Nguyen et al., 2020; Leal-Neto, Santos, Lee, Albuquerque, & Souza, 2020; Heidari et al., 2020; Bharati, Podder, & Mondal, 2020) have demonstrated the potential of AI to allow for rapid diagnosis of diseases. In particular, the use of AI to process medial images (Mansilla et al., 2020), and especially to detect TB from CXR images (Russakovsky et al., 2019) has made significant progress as models built on top of Deep Learning algorithms earn increasingly accurate results (Hwang, Kim, & Kim, 2016; Harris et al., 2019; Ahsan, Gomes, & Denton, 2019; Zeng et al., 2019).

There have been various studies that deal with the recognition of lung cancer (Han et al., 2019; Toğaçar et al., 2020). For instance, a recent work (Han et al., 2019) has been developed based on hybrid resampling and multi-feature fusion strategies. In that approach, hybrid resampling is used to minimize the risk of missing or large cavities, while multi-feature fusion strategies attempt to reserve context information of multiply CT windows more compactly. While the approach obtained an encouraging performance, it was experimented on a considerably small dataset. Similarly, an automated approach to detection of lung cancer has been built based on three deep learning models, including LeNet, AlexNet and VGG-16 (Toğaçar et al., 2020). To evaluate the resulting system, various experiments were conducted on an open dataset composed of Computed Tomography (CT) images. A recent work on classification of TB from CXR images (Rahman et al., 2020) employs various deep neural networks. By carefully investigating a wide range of related studies, we realized that while improvements have been achieved, there is still the need to further enhance the prediction accuracy as well as timing efficiency.

In this work, we build an expert system that can be deployed to serve the community at large. In particular, we propose a workable solution for the detection of TB from CXR images, exploiting cutting-edge Deep Learning techniques, namely EfficientNet (Tan et al., 2019) and Vision Transformer (Dosovitskiy et al., 2021). Moreover, we empower the learning process by incorporating various transfer learning strategies for EfficientNet, i.e., ImageNet (Rajpurkar et al., 2015), AdvProp (Whiting et al., 2019) and Noisy Student (Xie et al., 2020). An empirical evaluation on a considerably large dataset shows that our proposed models outperform some state-of-the-art studies.

In this respect, our paper has the following contributions:

- A system for recognition of TB from CXR images built on top of cutting-edge deep learning technologies.
- An empirical evaluation on a large and heterogeneous CXR image dataset.
- A software prototype provided as a mobile app is ready for download.[1]

The paper is structured as follows: Section 2 provides a brief description for convolutional neural networks of EfficientNet as well as three transfer learning methods, and Vision Transformer architecture. The dataset and metrics used for our evaluation are introduced in Section 3. Section 4 presents and analyzes the obtained results. We review the related work in Section 5 and conclude the paper in Section 6.

2. Background

This section first provides background on a family of deep neural networks, i.e., EfficientNet and Vision Transformer, which are used as the classification engine in our work. Afterwards, it gives a brief introduction to transfer learning as a base for further presentation.

In an attempt to improve the prediction accuracy, the EfficientNet deep neural network family (Tan et al., 2019) has been conceptualized by scaling a CNN in three dimensions, i.e., width, depth, and resolution. Among others, EfficientNet-B0 (shown in Fig. 1) is the most compact configuration with 18 convolution layers, each of them is convolved by a kernel of size (3,3) or (5,5). The next layers are decreased in resolution to reduce the feature map size, but increased in width, attempting to improve accuracy. For example, the second convolution layer consists of $W = 16$ filters, and the next convolution layer is with $W = 24$. The maximum number of filters is $D = 1{,}280$ by the last layer, which is fed to the final fully connected layer. The other configurations of the EfficientNet family are generated from EfficientNet-B0 by means of different scaling values (Tan et al., 2019).

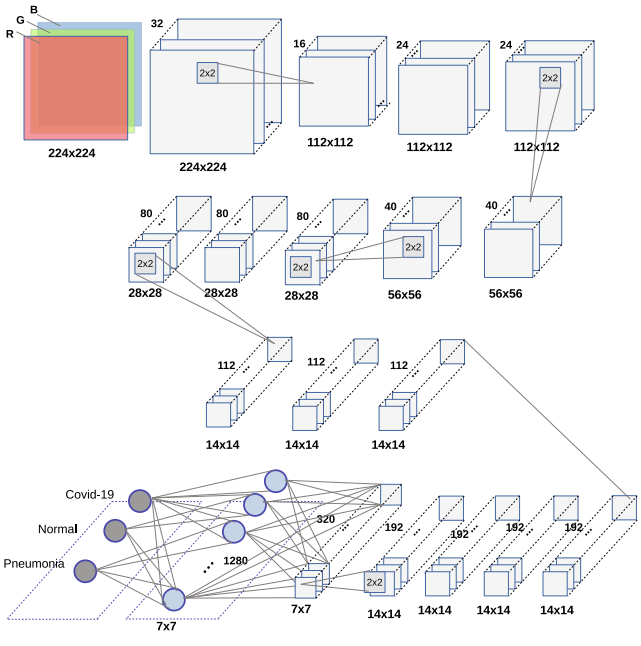


Fig. 1. The EfficientNet-B0 architecture.

[1] https://github.com/linhduongtuan/Tuberculosis_ChestXray_Classifier

Fig. 2 illustrates a hybrid model that uses the EfficientNet backbone to learn a 2D feature map of an input image. The model flattens at stage 4 into a sequence and supplements with a positional encoding before feeding images into a transformer encoder. Such a component then uses as input a small fixed number of learned positional embedding, which are called category queries, and repeatedly attends to the encoder output. We also add certain modified layers which work as a classifier for category *Tuberculosis* and *other Pneumonia*.

Recently, a novel network architecture, namely Transformer (Vaswani et al., 2017) has been proposed to take advantages of both Computer Vision and NLP architectures. Transformer is composed of an encoder-decoder structure, a self-attention mechanism, position-wise feed-forward layers with ReLU activations, Embedding and Softmax, and Positional Encoding. The encoder works as generations of a contextualized representation for every pixel channel in the input data, while the decoder corresponds to an autoregressive generation one channel per pixel at each time step. The self-attention mechanism is like a mapping of a query and a set of key-value pairs to an output. Transformer only uses self-attention instead of using sequence aligned RNNs or CNNs. Still, the deployment of transformer is in its infancy, though it has been widely used in NLP tasks.

Various attempts have been made to apply Transformer to image classification tasks by using a hybrid of self-attention and CNN architectures (Carion, Massa, Synnaeve, Usunier, & Kirillov, 2020) and replacing the entire convolutions (Wang et al., 2020). The Vision Transformer architectures have been conceived (Dosovitskiy et al., 2021) to follow the original architecture of Transformer for NLP (Vaswani et al., 2017). An image with height, width, and channel is reshaped into flattened 2D patches and used as the input for Transformer. In this respect, image patches resemble tokens in NLP implementations. The output of a Transformer encoder can be considered as the image presentation conducted by perpending a learnable embedding to the sequence of embedded patches. Furthermore, the positional information is retained by appending the position embeddings to the patch embeddings.

To optimize the computation, normally input images are scaled to $384 \times 384 \times 3$ in size, and the patch sizes are set to 14×14, 16×16, and 32×32 (Dosovitskiy et al., 2021). The sequence length input for Transformer is computed as $image\_resolution^2/patch\_size^2$. This aims at maintaining a trade-off between the computational expense and effectiveness.

In various NLP tasks, a popular way to train Transformer-bases models is to pre-train them on large corpora and then fine-tune them to meet the requirements of a specific task. Similarly, Vision Transformer architectures have been trained on the JFT-300 M dataset, consisting of 18 K classes with 300 million images (Sun, Shrivastava, Singh, & Gupta, 2017), and then the pre-trained weights are used for certain image recognition benchmarks. Such models yield promising accuracies, i.e., 88.55%, 90.77%, 94.55%, and 77.16% on ImageNet, ImageNet-ReaL, CIFAR-100, and VTAB, respectively (Dosovitskiy et al., 2021). Especially, the ViT-Huge model with 32 layers, hidden size of 1,280, multi layer perception of 5,120, attention heads of 16, and patch size of 14 earns an accuracy of 88.55% on the ImageNet dataset. In this respect, ViT-Huge outperforms the current state-of-the-art EfficientNet-L2 which gets an accuracy of 88.5% (Touvron et al., 2020) on the ImageNet dataset using pre-trained weights from Noisy Student (Xie et al., 2020). Concerning timing efficiency, ViT-Huge with image size of $384 \times 384 \times 3$ and patch size of 14 consumes 2,500 TPU days, while EfficientNet -L2 with image size of $600 \times 600 \times 3$ needs 12,300 TPU days (Dosovitskiy et al., 2021). This means ViT-Huge is approximately five times more efficient than EfficientNet-L2 with Noisy Student.

Normally, a CNN needs a huge amount of labeled data to train its internal weights and biases. Furthermore, a deeper network contains more parameters and thus it requires more data to avoid overfitting and to be effective. In this sense, it is crucial to train CNNs with *enough* data. However, such a constraint is difficult to be satisfied in the field, as the labeling process is normally done by hands, i.e., with the involvement of humans. Such a process is both time consuming and susceptible to error. To this end, transfer learning has been proposed as an effective method to excerpt and transfer the knowledge from a well-defined source domain to a novice target domain (Weiss, Khoshgoftaar, & Wang, 2016; Toğaçar et al., 2005). To be more concrete, transfer learning allows one to adopt existing well-trained convolution weights from a model which has been trained by means of large datasets. As it has been shown in various studies (Huang, Pan, & Lei, 2017; Duong, Nguyen, Di Sipio, & Di Ruscio, 2020), transfer learning is still beneficial to the target domain, even when the domain is quite different from the one where the original weights are borrowed. In the scope of our work, the following transfer learning strategies are considered:

- **ImageNet** (Rajpurkar et al., 2015): Weights trained through the ImageNet dataset have been used in several studies, as the dataset is made of more than 14 million images, spreading in miscellaneous categories;
- **AdvProp** (Whiting et al., 2019): Adversarial propagation has been proposed as an improved training scheme, with the ultimate aim of avoiding overfitting. The method treats adversarial examples as additional examples, and uses a separate auxiliary batch norm for adversarial examples;
- **NS** (Xie et al., 2020): This is an attempt to improve ImageNet classification Noisy Student Training by: *(i)* enlarging the trainee/student equal to or larger than the trainer/teacher, aiming to make the trainee learn better on a large dataset, and *(ii)* adding noise to the student, thus forcing it to learn more.

In this work, we develop an expert system which can be used to assist doctors in early detecting TB from CXR images. We build the system exploiting EfficientNet and Vision Transformer as the classification engines. Moreover, to accelerate the learning process, we propose obtaining network weights using the three different learning strategies for EfficientNet mentioned above, i.e., **ImageNet**, **AdvProp**, and **NS**. The succeeding section introduces the evaluation settings used to study the performance of our approach.

3. Evaluation

This section presents the research problems as well as the materials and methods used in the evaluation. Section 3.2 introduces three research questions. Afterwards, the dataset and experimental configurations are explained in Section 3.3 and Section 3.4, respectively. Finally, Section 3.5 lists all the evaluation metrics used to study the approach's performance.

We make use of recent implementations[2] of EfficientNet and Vision Transformer model, which have been built on top of the PyTorch framework.[3] Moreover, pre-trained weights have been imported from various sources to accelerate the training. The tool developed through this paper is made available in GitHub to facilitate future research.[4]

3.1. Architecture

The overall architecture is shown in Fig. 3. There are two main phases, namely training and deployment. The former starts with the data acquisition activities where training images are collected from different sources. Afterwards, expert radiologists are invited to carefully inspect and label the images based one their experience.

Images of size $320 \times 320 \times 3$ have been used for both the training and testing phases. To make the learning more effective, before training

[2] https://github.com/rwightman/pytorch-image-models
[3] https://pytorch.org
[4] https://github.com/linhduongtuan/Tuberculosis_ChestXray_Classifier

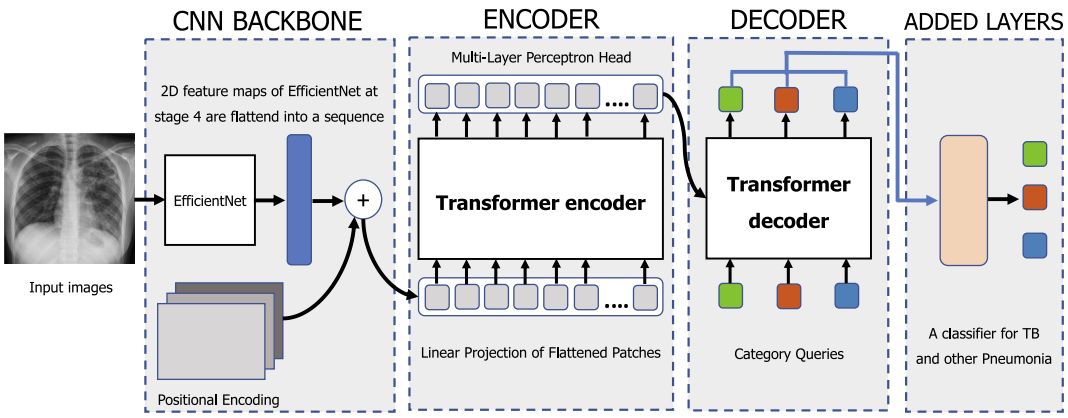


Fig. 2. EfficientNet backbone to learn 2D feature maps of an input image. The illustration of the hybrid model was reproduced from a recent work (Carion et al., 2020).

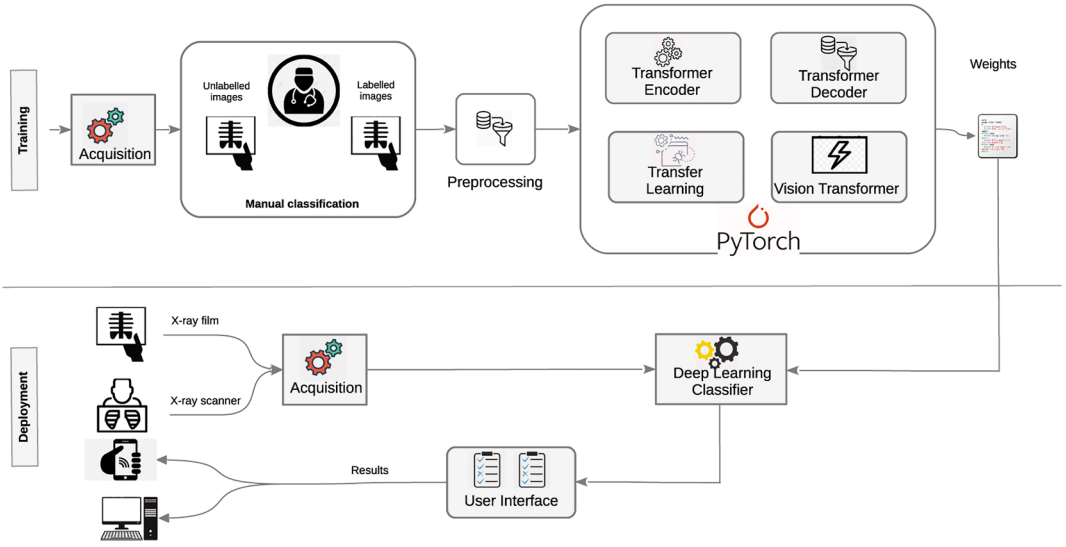


Fig. 3. The conceived architecture.

the deep neural networks, we performed various preprocessing steps on the input images. In particular, we employed augmentation strategies, which have been widely used in practice, so as to enrich the datasets, making them resemble real-world scenarios. There are the following steps: random rotation, random resized crop, random horizontal flip, color jitter, rand augment (Cubuk, Zoph, Shlens, & Le, 2020), and auto augment (Cubuk, Zoph, Mané, Vasudevan, & Le, 2019).

In reality, images can be heterogeneous, and such preprocessing phases aim to augment the dataset, attempting to cover real-world materials. The classified images are then used to train the system which runs on top of PyTorch and consists of different modules, namely Transformer Encoder, Transformer Decoder, Transfer Learning, and Vision Transformer. The final result of the training process is a set of internal parameters which are then applied to predict a label for input images. In the validation and testing phases, various weaker augmentation methods have also been applied, including resize, center crop, and color jitter.

By the testing/deployment phase, doctors or users can scan X-ray images with their PC, or upload them by mobile phones using the **Acquisition** module. The images are then fed as input for the detection engine which is the model learned from training data. The final prediction results are sent back to the doctors/users once the engine has performed the recognition phase.

By an empirical evaluation, we realized that by training with original neural network configurations, we obtained an adequate prediction accuracy. Aiming to increase the performance of our proposed classifiers for the two categories TB and other Pneumonia, we performed some modifications on the original framework as follows. We amended the last layer of the EfficientNet and Vision Transformer models by augmenting it with certain operators (see Table 1). In particular, we replaced the last Softmax layer with backbone models and interposed

Table 1
Proposed architecture based on EfficientNet and Vision Transformer families (N/A: not applicable).

Layer	Operator	Parameters
1	Linear	2,048 → 1,000
2	BatchNorm2D	1,000
3	Dropout	0.7
4	Linear	1,000 → 512
5	BatchNorm2D	512
6	Swish	N/A
7	Dropout	0.5
8	Linear	512 → 128
9	BatchNorm2D	128
10	Swish	N/A
11	Linear	128 → number of classes
12	Softmax	1

certain layers among the four linear, three BatchNorm2D, and two non-linearity activation layers, so called Swish. Moreover, to prevent overfitting, two Dropout layers are appended before the second and the third linear operators. The last linear layer is adjusted from 128 to the number of classes in the dataset. Eventually, we padded Softmax to the last linear layer, aiming to get a certain probability for each predicted image among the categories.

3.2. Research questions

In the scope of this paper, we are interested in understanding if the proposed architecture is able to provide accurate predictions in reasonable time frame for a real-world dataset. Thus we consider four research questions as follows:

- **RQ$_1$**: *Which network family yields the best prediction performance?* First, by means of a real-world dataset, we identify which neural network configuration obtains the best prediction performance.
- **RQ$_2$**: *What is the most suitable transfer learning strategy?* We imported pre-trained weights from three different transfer learning strategies, i.e., **ImageNet**, **AdvProp**, and **NS** to investigate which one between them is the most effective transfer learning method on the given dataset.
- **RQ$_3$**: *How does the proposed classifier compare with the baselines?* Considering two recent studies (Pasa et al., 2019; Souid et al., 2021) as baselines, we compare our approach with them to see how it performs with respect to effectiveness and efficiency.
- **RQ$_4$**: *How efficient are the models in predicting results?* Finally, we measure the average recognition speed to see if the proposed model is feasible in practice concerning timing efficiency.

3.3. Datasets

We populated a dataset from different publicly available data repositories as follows:

- The Montgomery County (MC) CXR dataset contains 138 frontal chest X-rays from Montgomery County's Tuberculosis screening program, of which 80 images are classified as normal cases and 58 images are with manifestations of TB (Jaeger et al., 2014);
- The Shenzhen dataset consists of 662 frontal CXR images, in which 326 images are classified as normal and 336 films are classified as manifestations of TB (Jaeger et al., 2014);
- The Belarus dataset is made of 304 CXR images of patients with confirmed TB;
- A Covid-19 dataset[5] compromises of 10 frontal CXR from TB cases (6/16 images were skipped because of duplication) and CXR images of other pneumonia (Cohen et al., 2020);
- Finally, we adopt additional images coming from the following sources: RSNA Pneumonia Detection Challenge dataset[6] (Wang et al., 2017); Covid-19 Radiography Database[7] (Carion et al., 2020) CXR images for categories normal and pneumonia caused by other pathogens.

From these repositories, we merged and removed any duplicate and/or corrupted images, and populated a dataset consisting of 28,672 images. Table 2 gives a summary for the resulting dataset. In particular, there are 28,840 CXR images in total, and *Pneumonia* is the largest category with 14,166 images. The *Normal* category consists of 13,808 images. The smallest category is *Tuberculosis* with only 698 images. Some examples of CXR images extracted from the dataset for different categories are shown in Fig. 4. Afterwards, the dataset was randomly split into training, validation, and testing sets, using the following ratio 80%, 10% and 10%, respectively. The training and validation sets are used to train as well as to tune our models to get the best weights. Afterwards, the obtained weights and biases are used to perform predictions on the test set.5.

Table 2
Chest X-ray tuberculosis dataset.

Type	Categories			Total
	Normal	Pneumonia	Tuberculosis	
Train	10,258	10,499	558	21,315
Validation	1,770	1,837	70	3,677
Test	1,780	1,830	70	3,680
Total	**13,808**	**14,166**	**698**	**28,672**

3.4. Settings

Deep neural networks such as EfficientNet and Vision Transformer require a platform with high computational performance. To conduct the evaluation, we used a server with hardware and software configurations listed in Table 3.

By means of an empirical evaluation, we realized that EfficientNet-B0 and EfficientNet-B1 are the most effective network configurations, compared to the other EfficientNet counterparts, thus we selected them for the final evaluation. Moreover, pre-trained weights for EfficientNet are obtained following all the transfer learning techniques mentioned in Section 2. Meanwhile, by the Vision Transformer family, there are six different configurations. First, we consider the following definitions:

- An image is divided into fixed-size patches. If the patch size is set as 16×16, then the final dimensions of the image is 48×48;
- Layers: They are a combination of a multi-head self-attention mechanism and position-wise fully connected feed-forward neural network;
- MLP: This is a class of feed-forward artificial neural network, and it consists of at least three layers of nonlinear activation nodes. Layers are named as an input and an output with one or more hidden layers;
- Hidden layers: They are placed between the input and output of the algorithm, in which the function uses weights to the inputs and passes them through an activation function as the output;
- Multi-heads attention: This is a map of a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors.

Due to the limited computational capability of the available server (cf. Table 3), we made use of an existing implementation of Vision Transformer models,[8] which offers some derivations from the original architectures, namely ViT_Small_Patch16_224 and ViT_Base_Patch16_224, which are based on the ViT_Base_Patch16_384 model. In particular, the input image size is scaled down from $384 \times 384 \times 3$ to $224 \times 224 \times 3$, and the other parameters such as layer numbers, head numbers and MLP size are also customized to construct ViT_Small_Patch16_224. The reduced models are more suitable for the available system.

Furthermore, instead of re-implementing hybrid models of ResNet (He, Zhang, Ren, & Sun, 2016) and Vision Transformer, we built hybrid models between EfficientNet and Vision Transformer. To be concrete, input images are passed into EfficientNet using pre-trained weight from ImageNet and extracted 2D feature outputs at stage 4. The sequenced feature maps are projected into Vision Transformer and then propagated to our modified layers which work as a classification engine for the

[5] https://github.com/ieee8023/covid-chestxray-dataset
[6] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[7] https://kaggle.com/tawsifurrahman/covid19-radiography-database
[8] https://github.com/rwightman/pytorch-image-models

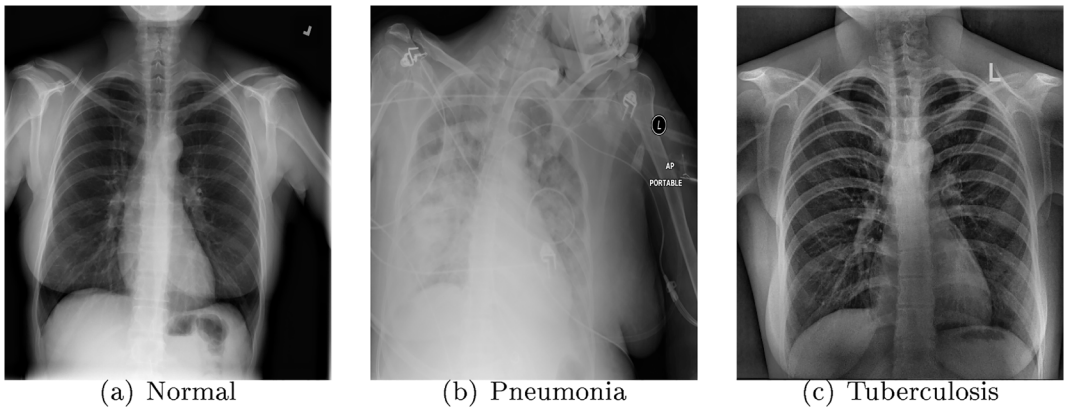


Fig. 4. Examples of CXR images.

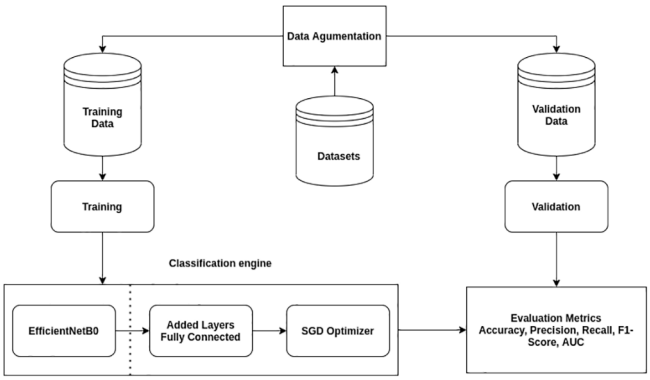


Fig. 5. Evaluation process.

Table 3
Hardware and software configurations.

Name	Description
RAM	96 GB
CPU	Intel®Xeon CPU E5-2678 V3 @ 2.50 GHz × 24
GPU	NVIDIA GeForce GTX 1080Ti
OS	Ubuntu 20.04
Python	3.7.5
Pytorch	1.6
Torchvision	0.7.0
Numpy	1.15.4
Timm	0.3.1

Tuberculosis and other *Pneumonia* categories.

For the hybrid Vision Transformer models since there are no fully pre-trained weights available, we cannot apply any transfer learning methods. In contrast, for ViT_Small_Patch16_224 and ViT_Base_-Patch16_224 we used pre-trained weights from ImageNet.

Combining the considered networks with three learning strategies mentioned in Section 2 results in 12 experimental configurations, i.e., C_i, i=1, .., 12 as shown in Table 4. *Batch size* corresponds to the number of items used for each training step; *# of Params* specifies the number of parameters used by each network; and finally *Size* is the file size needed to store the parameters. ViT_Base_EfficientNet_B1_224 is the largest network with respect to the number of parameters as well as the file size to store them. In particular, there are more than 89 millions of parameters which account for 703 MB of disk storage. The other Vision Transformer configurations are also large in size. For instance, the Vision Transformer models $C_7 \div C_{12}$ are made of more than 50 millions of parameters, resulting in more than 400 MB of trained weight size.

3.5. Evaluation Metrics

In the collected dataset, each image has been manually classified into one the categories, i.e., either *Normal* or *Pneumonia* or *Tuberculosis*, and we call them $G = (G_1, G_2, G_3)$, i.e., the ground-truth data. By running the classifiers on a test set, we obtained a label for each testing image, resulting in three predicted classes, i.e., $P = (P_1, P_2, P_3)$. The classification performance is measured by evaluating the relevance between the ground-truth labels G and the predicted ones P. Four metrics, namely *accuracy*, *precision* and *recall*, F_1 *score*, and Receiver Operating Characteristic (ROC) are used to measure the performance (Duong et al., 2020; Nguyen et al., 2021).

First, we consider the following definitions:

- A true positive is counted when the predicted label of an image matches with its real label; *TP* is the number of true positives, i.e., $TP_i = |G_i \cap P_i|, i = 1, 2, 3$;
- A false positive happens when the predicted label does not match with the real one; *FP* is the number of false positives, i.e., $FP_i = |P_i \setminus G_i|$;
- A false negative corresponds to an image that belongs to the ground-truth data but it is not classified, i.e., $FN_i = |G_i \setminus P_i|$;
- A true negative is an image that does not belong to the ground-truth data and it is not classified, i.e., $TN_i = \sum_{d \in C, d \neq c} |G_d|$.

Then the considered metrics are defined as follows:

Accuracy: Given a test set, accuracy is the fraction of correctly classified items to the total number of images.

$$accuracy = \frac{\sum_{i}^{3} TP_i}{\sum_{i}^{3} |G_i|} \times 100\% \quad (1)$$

Precision and Recall: Given a category, precision measures the ratio of the number of correctly classified images (true positives) to the total number of classified images (true positives plus false positives). Meanwhile, Recall is the ratio of the number of correctly classified items to the category (true positives) to the total number of items in the ground-truth data (true positives plus false negatives).

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

F_1 **score (F-Measure):** It is an average of precision and recall scores:

Table 4
Experimental configurations.

Configuration	Network	Patch size	# of layers	Hidden size	MLP size	# of heads	Image size	Batch size	# of Params	Transfer Learning	Size (MB)
C_1	EfficientNet-B0	—	18	—	—	—	320 × 320 × 3	50	7,919,391	ImageNet	53.1
C_2	EfficientNet-B0	—	18	—	—	—	320 × 320 × 3	50	7,919,391	AdvProp	53.1
C_3	EfficientNet-B0	—	18	—	—	—	320 × 320 × 3	50	7,919,391	Noisy Student	53.1
C_4	EfficientNet-B1	—	20	—	—	—	240 × 240 × 3	70	10,424,898	ImageNet	73.4
C_5	EfficientNet-B1	—	20	—	—	—	240 × 240 × 3	70	10,424,898	AdvProp	73.4
C_6	EfficientNet-B1	—	20	—	—	—	240 × 240 × 3	70	10,424,898	Noisy Student	73.4
C_7	ViT_Small_Eff_B0_224	16	26	768	2,304	8	224 × 224 × 3	80	53,754679	None	426.1
C_8	ViT_Small_Eff_B1_224	16	28	768	2,304	8	224 × 224 × 3	60	56,213,467	None	446.3
C_9	ViT_Base_Eff_B0_224	16	30	768	3,072	12	224 × 224 × 3	60	92,309,279	None	728.4
C_{10}	ViT_Base_Eff_B1_224	16	32	768	3,072	12	224 × 224 × 3	50	94,814,915	None	748.6
C_{11}	ViT_Small_Patch16_224	16	8	768	2,304	8	224 × 224 × 3	50	51,385,251	ImageNet	400.6
C_{12}	ViT_Base_Patch16_224	16	12	768	3,072	12	224 × 224 × 3	50	89,170,851	ImageNet	703
C_{13}	Convolutional Neural Network	—	22	—	—	—	512 × 512 × 1	512	231,203	None	2.0
C_{14}	MobileNetV2	—	19	—	—	—	224 × 224 × 3	70	6,135,715	ImageNet	38.9

$$F_1 = 2 \cdot \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{4}$$

False Positive Rate (FPR): It measures the fraction of the number of images being falsely classified (false positives), to the total number of images being either correctly not classified, or falsely classified (false positives plus true negatives).

$$FPR_i = \frac{FP_i}{FP_i + TN_i} \tag{5}$$

True Positive Rate (TPR): It is equal to Recall:

$$TPR_i = Recall_i \tag{6}$$

Receiver Operating Characteristic (ROC): In a 2D space, where the x-Axis represents FRP and the y-Axis corresponds to TPR, a receiver operating characteristic (ROC) depicts the relationships between FPR and TPR (Fawcett, 2006). An ROC curve located close to the upper left corner corresponds to a better prediction performance than the one that resides at the lower right corner in the 2D space. The area under the ROC curve (AUC) is an explicit indication of how good a classifier is. An AUC value of 0.5 corresponds to a random guessing, while the AUC value of 1.0 corresponds to a perfect classification.

Speed: Besides accuracy, we also pay our attention to efficiency, considering the fact that the model needs to return a final prediction in a reasonable amount of time. The system presented in Table 3 is used to benchmark the processing time, i.e., the average number of predicted items in a second.

4. Results

This section presents in detail the experimental results by referring to the research questions introduced in Section 3.2.

4.1. RQ_1: *Which network family yields the best prediction performance?*

As shown in Table 4, we experimented with various configurations concerning networks and transfer learning techniques. This research question aims to find the neural network configuration that brings in the best accuracy. This is important in practice as it helps us choose the most suitable setting when running the framework on real data.

We report the precision, recall and F_1 scores in Table 5. Among the considered configurations, the classifier achieves the best performance with C_{10}, i.e., accuracy is 97.72%. That means ViT_Base_Eff_B1_224 helps obtain the best accuracy. In the second place, EfficientNet-B1 trained with weights from Noisy Student, i.e., C_6, gets 97.61%.

With respect to precision, both C_5 and C_{10} achieve 0.973 for the *Normal* category, being the configurations with the best precision. Meanwhile, C_9 and C_{10} get the maximum precision for the TB category. Similarly, when considering F_1 score, we can see that C_{10} earns the best performance on the *Normal* and *Pneumonia* categories.

The experiments confirm that Vision Transformer is beneficial as it helps obtain a superior performance on different metrics. Compared to the original EfficientNet family, the Vision Transformer architectures, i.e., $C_7 \div C_{12}$, generally bring a better performance.

Fig. 6 and Fig. 7 depict the ROC curves for all the considered configurations for the three categories, i.e., 0 - *Normal*, 1 - *Pneumonia*, and 2 - *Tuberculosis*. For the EfficientNet configurations, i.e., $C_1 \div C_6$, we can see that the classifier obtains a high AUC, i.e., the minimum value is 0.97 and the maximum value is 1.00. By the Vision Transformer configurations, i.e., $C_7 \div C_{12}$, a clear improvement compared to the EfficientNet configurations is seen, by all the categories, the minimum AUC is 0.99, and the maximum AUC is 1.00.

Altogether, it is clear that the Vision Transformer models enable the original architecture to achieve a performance gain. In this respect, we come to conclusion that the application of Vision Transformer is beneficial to the recognition of TB from CXR images as it helps boost up the accuracy considerably.

Answer to RQ_1: Although both network families obtain high accuracy and precision, the classification using Vision Transformer yields the best prediction performance.

4.2. RQ_2: *What is the most suitable transfer learning strategy?*

Transfer learning has been widely used to make the training more

Table 5
Experimental results.

Configuration		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}
Accuracy (%)		97.39	97.47	96.82	96.44	97.58	97.61	97.26	96.58	97.34	97.72	95.92	96.17	96.25	96.47
Precision	Normal	0.968	0.972	0.957	0.948	0.973	0.968	0.965	0.953	0.967	0.973	0.946	0.947	0.976	0.958
	Pneu.	0.981	0.977	0.978	0.981	0.978	0.983	0.980	0.978	0.978	0.982	0.971	0.977	0.950	0.963
	TB	0.959	0.986	0.972	0.972	0.986	0.986	0.986	0.971	1.000	1.000	0.986	0.959	0.969	0.993
Recall	Normal	0.979	0.977	0.978	0.981	0.978	0.983	0.979	0.979	0.978	0.982	0.971	0.976	0.948	0.963
	Pneu.	0.968	0.971	0.957	0.947	0.972	0.968	0.965	0.953	0.968	0.972	0.947	0.946	0.979	0.957
	TB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.957	1.000	0.985	0.971	1.000	0.900	1.000
F_1-**score**	Normal	0.973	0.974	0.968	0.964	0.975	0.976	0.972	0.966	0.972	0.976	0.959	0.961	0.962	0.961
	Pneu.	0.974	0.974	0.967	0.963	0.975	0.975	0.972	0.965	0.973	0.977	0.959	0.961	0.964	0.960
	TB	0.979	0.993	0.986	0.986	0.993	0.993	0.993	0.964	1.000	0.993	0.978	0.979	0.933	0.996

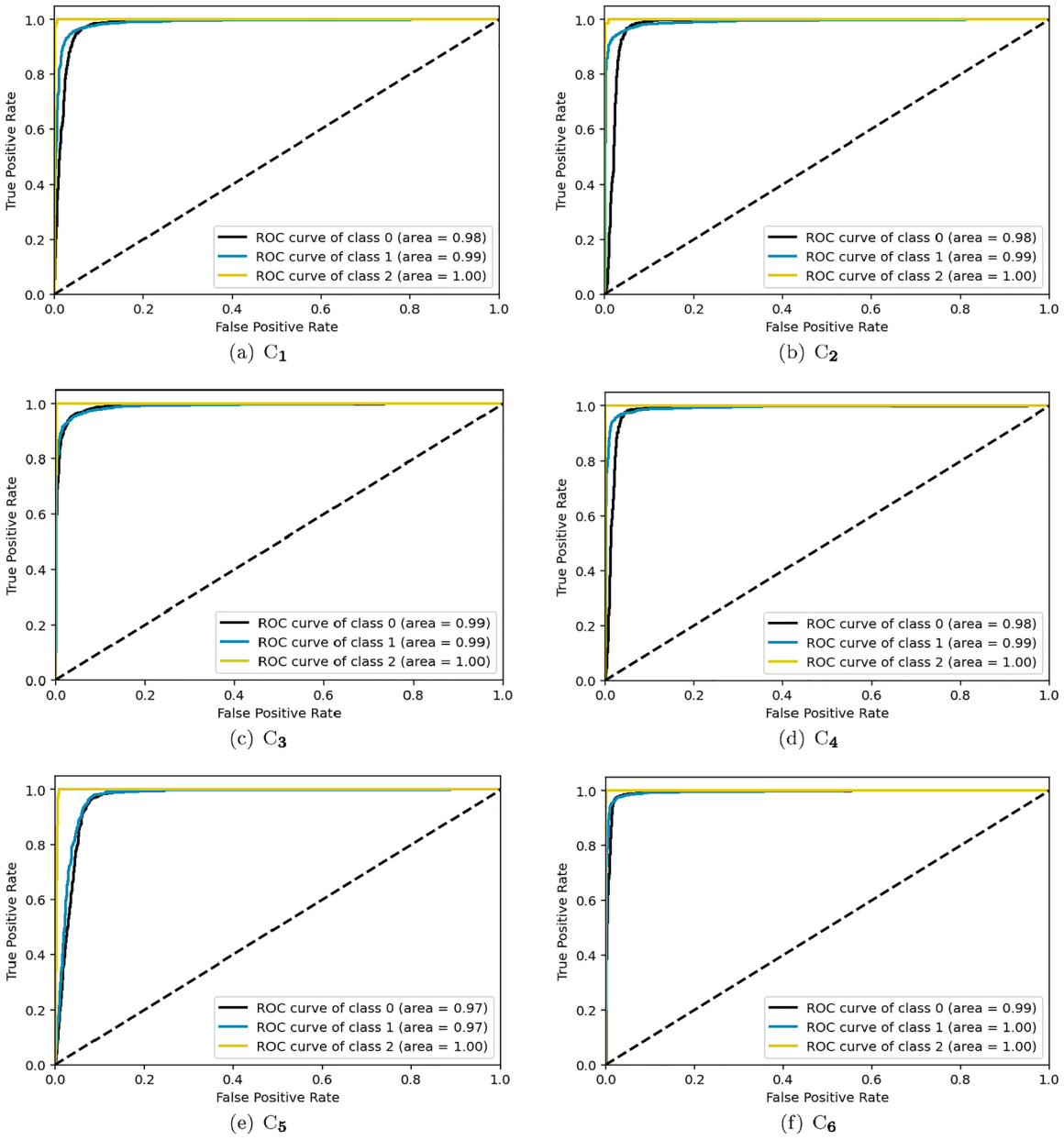


Fig. 6. Receiver operating curves of EfficientNet for $C_1 \div C_6$.

effective as well as efficient (Huang et al., 2017). There are various learning strategies, and among others, **ImageNet**, **AdvProp**, and **NS** (see Section 2) have been selected for our experiments as they have demonstrated their effectiveness in various applications (Rajpurkar et al., 2015; Whiting et al., 2019; Xie et al., 2020). In this research question, we want to find out which one among these techniques is more suitable for the detection of TB on the considered datasets.

Fig. 8 and Fig. 9 report the confusion matrices for all the

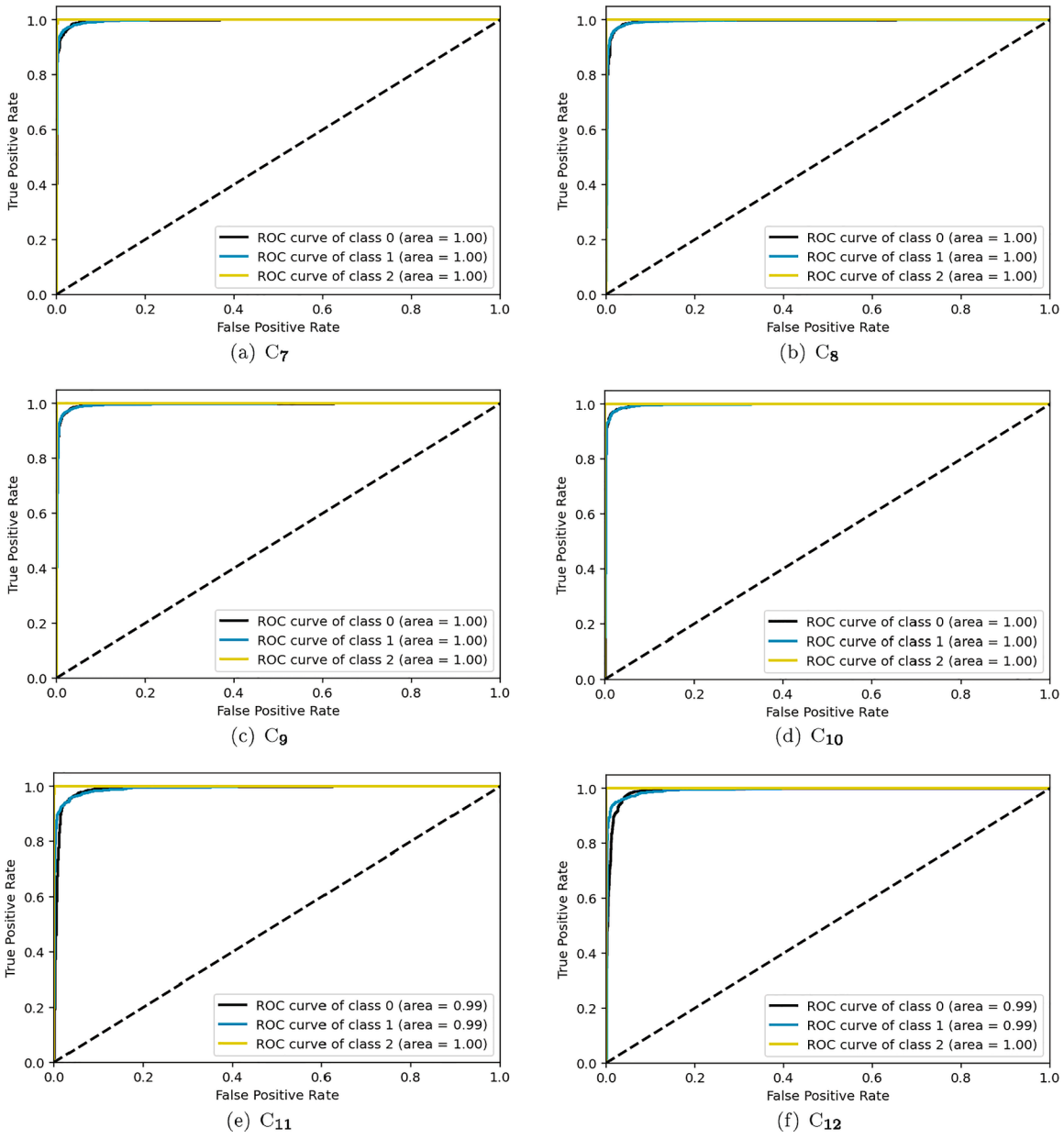


Fig. 7. Receiver operating curves of Vision Transformer for $C_7 \div C_{12}$.

configurations. We see that each transfer learning technique has a different effect on the categories as explained below.

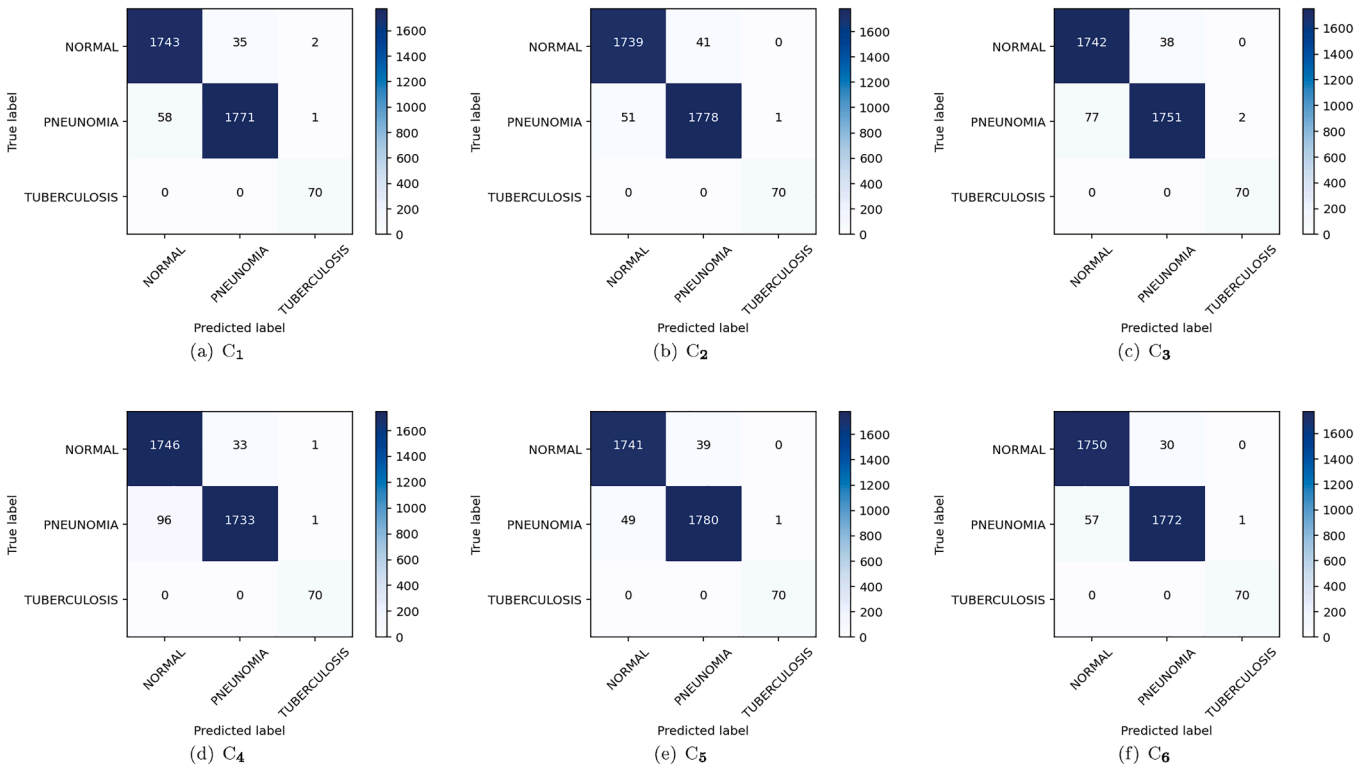
In particular, concerning the *Normal* category, C_6 is the configuration that brings the best performance, i.e., 1,750 among 1,780 images are correctly classified. That means, on the given dataset, using **Noisy Student** with EfficientNet-B1 as the classification engine is beneficial to the detection of the *Normal* category. Meanwhile, for the *Pneumonia* category, C_5 obtains the largest number of correctly classified images, i.e., 1,780 among 1,830 items are predicted with the correct label. The classifier also obtains a good performance by the C_{10} configuration, i.e., it correctly classifies 1,779 among 1,830 items, corresponding to 97.21%. Though this configuration corresponds to randomization, i.e., without applying any transfer learning technique, it uses the ViT_Base_Eff_B1_224 network. In this case, we come to the conclusion that **AdvProp** is beneficial to *Pneumonia*, given that EfficientNet -B1 is the network configuration.

On the other hand, we can see that the classifier obtains the maximum prediction performance for the *Tuberculosis* category by almost all the configurations. To be concrete, except C_8, C_{10} and C_{11}, the classifier correctly recognizes all the considered *Tuberculosis* images by the remaining configurations. **Answer to** RQ_2: All the transfer learning techniques are beneficial to the detection of *Tuberculosis* category. Meanwhile, Noisy Student and AdvProp are helpful for the detection of *Normal* and *Pneumonia*, respectively.

4.3. RQ_3: *How does the proposed classifier compare with the baselines?*

Both the considered studies (Pasa et al., 2019; Souid et al., 2021) obtained an encouraging prediction performance. Thus, in this research question we consider them as the baselines for comparison with our approach, corresponding to Configurations C_{13} and C_{14} (see Table 4). Nevertheless, by running the original code[9] of one of the baselines (Pasa et al., 2019) on the considered dataset (see Section 3.3), we encountered

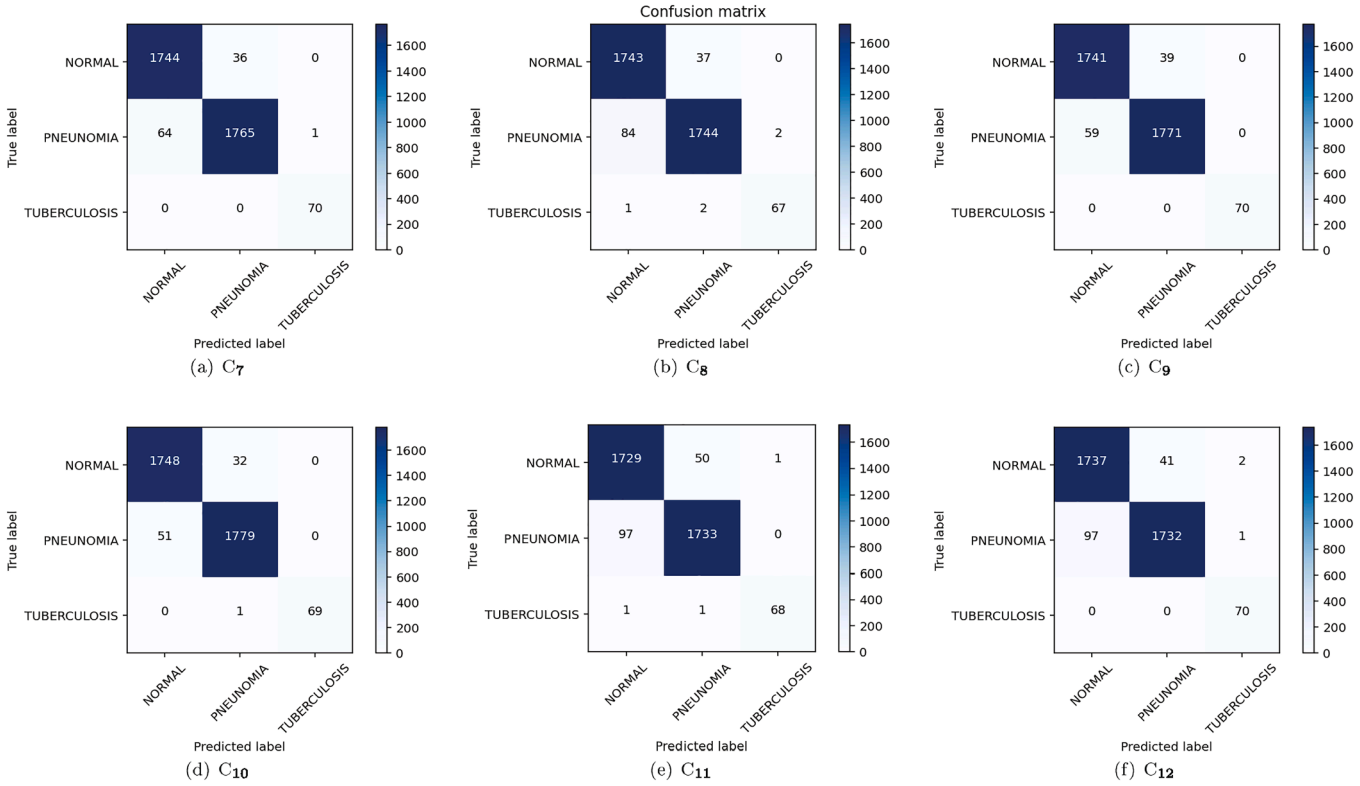
[9] https://github.com/frapa/tbcnn



(a) C_1 (b) C_2 (c) C_3

(d) C_4 (e) C_5 (f) C_6

Fig. 8. Confusion matrices for $C_1 \div C_6$.



(a) C_7 (b) C_8 (c) C_9

(d) C_{10} (e) C_{11} (f) C_{12}

Fig. 9. Confusion matrices for $C_7 \div C_{12}$.

the following problem. The original tool was implemented on TensorFlow 1 and it works with images of size 512×512. However, on a large dataset, the tool cannot load a large number of images, considering the fact that the dataset consists of more than 20 K images for training and 3 K images for testing. In particular, the tool has some issues with memory management and it exceeds 96 Gb of RAM when loading the dataset. Due to these reasons, we decided to re-implement the tool by strictly following the descriptions in the original paper (Pasa et al., 2019). We

only changed the batch size to optimize the memory usage. Eventually, we ran the code using 400 epochs on the given dataset and the training phase took 12 h to commit.

The final experimental results are shown in Fig. 10. Though the approaches obtain high AUC values (see Fig. 10a and Fig. 10b), as seen in Fig. 10c and Fig. 10d, it is clear that the prediction performance fluctuates between the three categories. In particular, by the first baseline (Pasa et al., 2019), for the *Normal* category, the tool correctly classifies 1,687 over 1,780 images, corresponding to an accuracy of 94.77%. For the *Pneumonia* category, 1,792 among 1,830 are correctly classified, resulting in an accuracy of 97.92%. Finally, for the *Tuberculosis*, 63 among 70 images are assigned to their original category.

Compared to the baselines, our approach obtains a much better performance, by almost all the configurations. Referring back to the results in Fig. 8 and Fig. 9, we can see that our proposed tool outperforms by the *Normal* and *Tuberculosis* categories. For example, C_6 obtains the best performance for the *Normal* category, i.e., 1,750 among 1,780 images are correctly classified. Meanwhile, the baseline gets only 1,687 true positives for this category. Considering ROC in Fig. 10a and Fig. 10b, we also see that the AUC values for the categories obtained by the baselines are lower than the other configurations. It is evident that the approach presented in this paper earns a superior prediction performance compared to that of the two baselines. **Answer to** RQ_3: Our proposed approach outperforms both baselines with respect to accuracy and ROC.

4.4. RQ_4: *How efficient are the models in predicting results?*

For this research question, we measure the time needed to perform predictions on the considered dataset, i.e., the average number of predicted items in a second. This aims at investigating the feasibility of our proposed approach in practice, i.e., if it can return predictions in a timely manner. We use the system specified in Table 3 to measure the processing time.

In Fig. 11, we sketch the recognition speed for all the considered configurations. Most of the them have a comparable execution time on the dataset to produce the final results. For instance, six among 12 configurations, i.e., C_1,C_2,C_6,C_7,C_9, and C_{11} return 147 predictions per

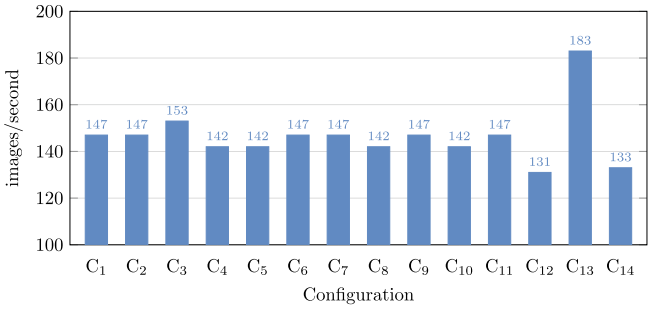
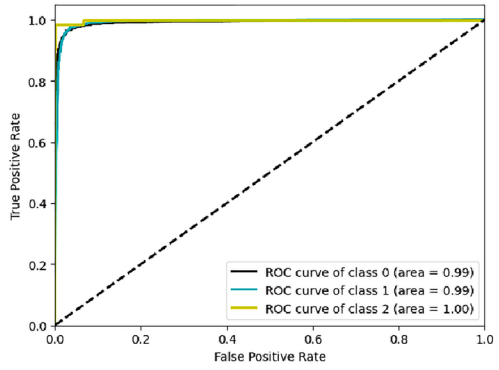
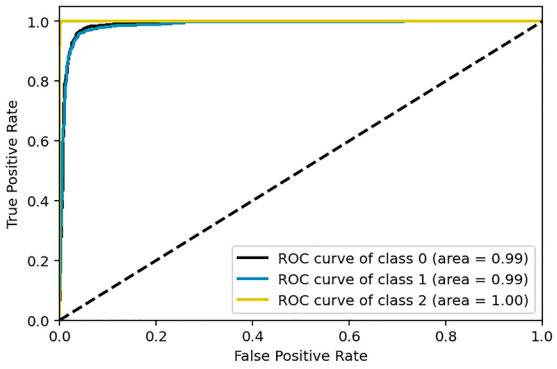


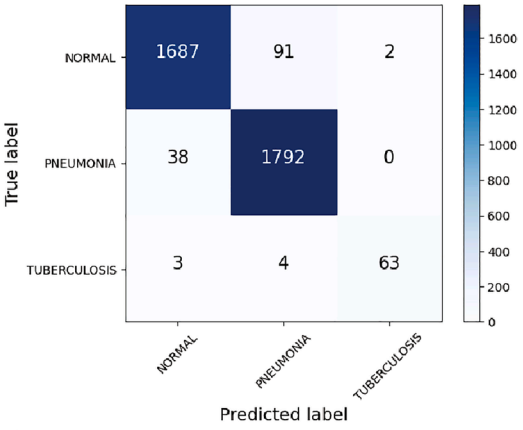
Fig. 11. Recognition speed for the configurations on the considered dataset.



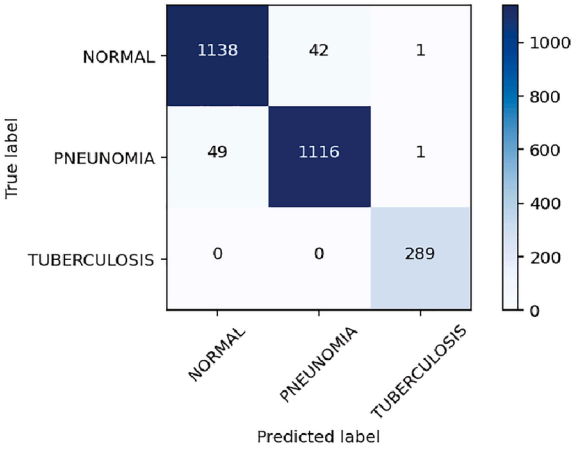
(a) ROC of the work by Pasa *et al.* (Pasa et al., 2019)



(b) ROC of the work by Souid *et al.* (Souid et al., 2021)



(c) Confusion matrix of the work by Pasa *et al.* (Pasa et al., 2019)



(d) Confusion matrix of the work by Souid *et al.* (Souid et al., 2021)

Fig. 10. Performance obtained by the baselines.

second; three of them return 142 predictions per second, i.e., i.e., C_4,C_5, and C_8. The most efficient configuration is C_3, i.e., EfficientNet-B0 trained with Noisy Student weights. The most time consuming configuration is C_{12}, or ViT_Base_Patch16_224 with pre-trained weights from ImageNet. The first baseline (Pasa et al., 2019), i.e., Configuration C_{13} yields a better efficiency compared to our approach, as it returns 183 images per second. This is understandable as the network is quite compact and it needs a smaller number of parameters. Together with the results from RQ_1 and RQ_2, we can see that the most effective configurations are not necessarily the most timing efficient ones. In particular, it is clear that while the configurations built with Vision Transformer are the least efficient ones, i.e., they are more time consuming compared to the remaining settings. **Answer to** RQ_4: EfficientNet-B1 trained with ImageNet is the most efficient configuration. Though Vision Transformer obtains a better prediction accuracy, it suffers from a low timing efficiency.

4.5. Threats to Validity

We discuss the threats that may influence the internal, external, and conclusion validity of our findings.

Internal validity. This concerns the factors that might have an adverse effect on the final results. A probable threat might come from the results for the *Tuberculosis* category as it contains a small number images. Such a threat is partly eliminated by the other two categories in the dataset, i.e., *Normal* and *Pneumonia*, since they have a much larger number of images. Still, we believe that the threat can only be completely mitigated when there are more labeled images for the *Tuberculosis* category. We plan to re-evaluate the proposed approach when there is a bigger dataset available. Moreover, the re-implementation of the baseline (Pasa et al., 2019) may induce a threat to internal validity. We attempt to minimize the threat by strictly following the description in the original paper. We only optimize the memory usage to make the framework run on a larger dataset.

External validity. The is related to the factors that might negatively affect the generalizability of our findings for the scenarios outside the scope of this work, for instance, in practice there is a limited amount of training data available. We moderated this threat by splitting the dataset into train/validation/test parts. In this way, we simulate an actual use of the system where only one part of the data is fed to the system as input, while the remaining parts are used for calibrating and testing the model.

Conclusion validity. The evaluation metrics accuracy, precision, recall, F_1, ROC, and execution time may possibly yield a conclusion threat. To cope with this issue, we adopted such measures as recommended by the previous scientific literature related to our setting, and employed the same metrics for evaluating all the classifiers, including the baseline.

5. Related work

In recent years, we have witnessed a proliferation of Deep Learning across several application domains. This section reviews notable studies to the deployment of deep learning in the healthcare sector (Section 5.1). Afterwards, we recall related work in the domain of automatic recognition of tuberculosis from chest X-ray (CXR) images (Section 5.2).

5.1. Applications of deep learning in healthcare

A new computer-aided detection system has been proposed to classify benign and malignant mass tumors in mammograms using convolutional neural networks (Ragab, Sharkas, Marshall, & Ren, 2019). Two datasets being extracted from 9,368 images were used for evaluation. The approach obtained a maximum accuracy of 87.2% and AUC gain of 0.94. Agarwal, Diaz, Lladó, Yap, and Martí (2019) investigated the performance of VGG16, ResNet50, and InceptionV3 for classifying mass and non-mass breast regions for the various mammograms database. A dataset with 1,592 and 112 images from CBIS-DDSM and INbreast has been populated to use as the evaluation data. A total of 81,766 with a half of positive and the other of negative images are extracted, and the highest accuracy obtained by using InceptionV3 was 98%.

A deep learning model named COVNet (Li et al., 2009; Li et al., 2020) was developed to extract visual features from volumetric chest CT exams, aiming to detect COVID-19. The proposed model has been evaluated using community acquired pneumonia (CAP) and other non-pneumonia CT exams collected from different hospitals in more than three years. The approach's performance was studied by means of Precision, Recall, and AUC.

There have been various studies that deal with the recognition of lung cancer (Han et al., 2019; Toğaçar et al., 2020). For instance, a recent work (Han et al., 2019) has been developed based on hybrid resampling and multi-feature fusion strategies. In that approach, hybrid resampling is used to minimize the risk of missing or large cavities, while multi-feature fusion strategies attempt to reserve context information of multiply CT windows more compactly. While the approach obtained an encouraging performance, it was experimented on a considerably small dataset. Similarly, an automated approach to detection of lung cancer has been built based on three deep learning models, including LeNet, AlexNet and VGG-16 (Toğaçar et al., 2020). To evaluate the resulting system, various experiments were conducted on an open dataset composed of Computed Tomography (CT) images.

5.2. Automated detection of tuberculosis with deep learning

We review the most notable studies to deal with the recognition of TB from CXR images, paying attention to the dataset used and the final prediction accuracy. Table 6 provides a summary of the studies considered in this subsection.

Jaeger et al. (2014) evaluate various machine learning techniques to classify Tuberculosis CXR images on the MC (Jaeger et al., 2014) and Shenzhen datasets (cf. Section 3.3). The authors reported a maximum AUC of 0.87 and an accuracy of 78.3% on the MC dataset, and an AUC of 0.90, an accuracy 84% on the Shenzhen dataset. In comparison to the work by Jaeger et al. (2014), our proposed approach considerably improves the prediction performance, even on larger datasets.

Huang et al. (2016) performed an evaluation of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) on various datasets. The first one was obtained from the Korean Institute of Tuberculosis (KIT) with 10,848 CXR images, the second one consists of 662 CXR images from the Shenzhen dataset, and the third one with CXR 138 images collected from the National Institutes of Health (NIH). The obtained AUC values are as follows 0.964, 0.88, and 0.93 for the KIT, NIH, and Shenzhen datasets, respectively. In the present paper, with the application of Vision Transformer and different transfer learning strategies, we are able to improve the final accuracy.

The proposed model in a recent work (Pasa et al., 2019) is simple, yet very effective. In comparison to other existing studies (Huang et al., 2016; Lakhani & Sundaram, 2017), the model achieved a comparable

Table 6
A summary of the related studies on the detection of TB from CXR.

Approach	Data size	Technique	AUC	Accuracy
Jaeger et al. (2014)	662	SVM	0.9	0.84
Huang et al. (2016)	10,848	CNN	0.96	0.90
Lakhani and Sundaram (2017)	1,007	AlexNet, GoogLeNet	0.99	0.96
Pasa et al. (2019)	1,104	CNN	0.92	0.86
Vajda et al. (2018)	662	Neural Network	0.99	0.97
Rajpurkar et al. (2020)	677	DenseNet	0.78	0.79
Ahsan et al. (2019)	1,600	VGG16	–	0.81
Bharati et al. (2020)	5,606	Hybrid VGG	–	0.73

performance. The maximum accuracy and AUC on the dataset combined from the MC and Shenzhen datasets were 86.2% and 0.925, respectively. The study focused on detecting two categories, i.e., *Tuberculosis* and *non-TB*, and other pathology images on CXR, i.e., lung cancers, bacterial pneumonia, viral pneumonia, can be marked as positive. In our work, we re-implemented the baseline (Pasa et al., 2019) on top of the PyTorch framework. The experimental results show that our proposed clearly outperforms the existing framework.

Lakhani et al. leveraged two popular deep learning structures, namely GoogLeNet (Sutoko et al., 2015) and AlexNet (Krizhevsky et al., 2012) to classify images containing manifestations of pulmonary tuberculosis or healthy case (Lakhani & Sundaram, 2017). The dataset includes 1,007 CRX images with 68% being used for training, 17.1% for validation, and 14.9% for testing. The ensemble of the AlexNet and GoogLeNet gained an AUC of 0.99.

A recent study (Vajda et al., 2018) achieved the best ROC and accuracy of 97.03% and 0.99 on the Shenzhen dataset. Using the MC dataset, however, a lower accuracy and AUC were obtained, i.e., 84.75% and ROC, respectively. The performance of AI techniques based on software for the detection of abnormal regions related to pulmonary Tuberculosis (computer-aided detection, CAD) on CXR images has been recently investigated (Harris et al., 2019). The authors considered 4,712 studies including 40 development and 13 clinical studies. They also indicated that most of the pulmonary TB recognition was conducted based on development studies or CAD design methods. Furthermore, the authors evaluated the performance by Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 (Whiting et al., 2011), and the CAD design methods presented promising results in comparison with Clinical studies. Specifically, the median AUC of development studies was 0.88 whereas clinical studies gained 0.75 as AUC. The authors also demonstrated the potential source of bias due to the detection of TB on images by human reader instead of microbiological testing of sputum (Harris et al., 2019). A recent work on classification of TB from CXR images (Rahman et al., 2020) employs various deep neural networks. By carefully investigating a wide range of related studies, we realized that while improvements have been achieved, there is still the need to further enhance the prediction accuracy as well as timing efficiency.

Rajpurkar et al. (2020) present a deep learning approach to diagnose TB based on CXR images and clinical information. The CXR images are collected from 677 HIV patients who have symptom of TB. The dataset was and split into training and testing datasets, the former consists of 563 images for both positive and negative cases whereas the latter includes 114 images. The authors also applied clinical covariates to the algorithm, namely age, oxygen saturation, hemoglobin, CD4 count, white blood cell count, temperature, current antiretroviral therapy status, and the patient's prior history of TB. The classification task was carried out by using DenseNet (pre-trained on the CheXpert) to extract image features as a vector with high dimensionality. Then, the network is split into two modules, diagnosing of tuberculosis and predicting the occurrence of six findings. The diagnosis module uses 20 image features, 8 clinical covariates, and passes into a two-layer neural network to classify the disease. Furthermore, the findings module consists of a linear multi-label classifier utilized with the high dimensional vector with 6 output units. The proposed approach achieved an accuracy of 0.79 on the cases the physicians with algorithm assistance whereas the physicians obtain a mean accuracy of 0.65 in the same cases.

In this work, we attempt to fill the existing gap, aiming to improve the prediction performance on larger TB datasets. In fact, all the aforementioned studies have been tested with a limited amount of data, while in our work, we fused different data repositories to form a large dataset. Compared to these studies, by employing Vision Transformer and transfer learning, we achieve a better prediction performance with respect to different quality metrics. We also demonstrated that our proposed approach is both effective and efficient, it obtains a high prediction accuracy while maintaining a low response time. In conclusion, we suppose that the approach is feasible in practice, and ready to be deployed.

6. Conclusion and future work

This work presents a workable solution for the detection of Tuberculosis from chest X-ray images. Starting from the observation that while existing approaches obtained an encouraging prediction performance, most of them have been evaluated on small and undiverse datasets, we hypothesize that such a good performance might not hold for heterogeneous data sources, which originate from real world scenarios. Our model has been implemented based on two building blocks: deep convolutional neural networks with EfficientNet and *Attention* with Vision Transformer as the prediction engines, and effective transfer learning algorithms. One of the main advantages of EfficientNet is that the network family is compact as it is small in size and efficient, allowing us to incorporate various augmented techniques, e.g., Vision Transformer and Transfer Learning. An empirical evaluation on a considerably large dataset combined by using various datasets, which have been widely used in different papers, shows that our system obtains a better prediction performance compared to relevant studies. We conclude that the combination of EfficientNet with Vision Transformer and Learning brings in substantial improvement in performance compared to state-of-the-art approaches.

For our future work, we plan to incorporate more baselines for comparison with the conceived tool. This aims to better study the framework's intrinsic characteristics and highlight its novelty compared to existing ones. More importantly, we are now in the process of collaborating with various hospitals in Vietnam to collect additional data from radiologists, also by strictly following ethical guidelines for healthcare related research. We believe that a new tuberculosis dataset will help us further refine the learned model, and thus improving its effectiveness in real-world use cases. Finally, we also aim to provide doctors with a practical means to assist their daily work. To this end, we already started with the development of a prototype for the framework on the Android operating system. In the near future, we are going to launch a mobile app which can conveniently run on lightweight devices such as smartphones or tablets. The app is expected to work as a non-invasive pre-screening tool, being able to serve doctors at large. Such a tool is highly suitable for many mountainous and remote regions in Vietnam, where there is a lack of proper facilities to support healthcare specialists.

CRediT authorship contribution statement

Linh T. Duong: Conceptualization, Software, Writing - original draft, Writing - review & editing. **Nhi H. Le:** Software, Validation. **Toan B. Tran:** Validation, Visualization. **Vuong M. Ngo:** Writing - original draft. **Phuong T. Nguyen:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., & Martí, R. (2019). Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging, 6*(3), Article 031409.

Ahsan, M., Gomes, R., & Denton, A. (2019). Application of a convolutional neural network using transfer learning for tuberculosis detection. In *2019 IEEE International Conference on Electro Information Technology (EIT)* (pp. 427–433). https://doi.org/10.1109/EIT.2019.8833768

Alizadeh, R., Allen, J.K., Mistree, F., 2020. Managing computational complexity using surrogate models: a critical review. 31 (3): 275–298. ISSN 1435-6066. DOI: 10.1007/s00163-020-00336-7.

Alizadeh, R., Jia, L., Nellippallil, A. B., Wang, G., Hao, J., Allen, J. K., & Mistree, F. (2019). Ensemble of surrogates and cross-validation for rapid and accurate predictions using small data sets. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 33*(4), 484–501. https://doi.org/10.1017/S089006041900026X

Bharati, S., Podder, P., & Mondal, M.R.H. (2020). Hybrid deep learning for detecting lung diseases from x-ray images. Informatics in Medicine Unlocked, 20: 100391, 2020. ISSN 2352–9148. doi: 10.1016/j.imu.2020.100391. https://www.sciencedirect.com/science/article/pii/S2352914820300290.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, In Proceedings of the 16th ECCV 2020, volume 12346 of Lecture Notes in Computer Science, pages 213–229. Springer, 2020. DOI: 10.1007/978-3-030-58452-8_13. doi: 10.1007/978-3-030-58452-8_13.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B. I., & Islam, M. T. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access, 8*, 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287

Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., & Ghassemi, M. (2020). COVID-19 image data collection: Prospective predictions are the future. CoRR, abs/2006.11988, 2020. URL https://arxiv.org/abs/2006.11988.

Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., & Le, Q.V. (2019). Autoaugment: Learning augmentation strategies from data. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pages 113–123. Computer Vision Foundation/ IEEE, 2019. DOI: 10.1109/CVPR.2019.00020. http://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html.

Cubuk, E.D., Zoph, B., Shlens, J., & Le, Q., (2020). Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 18613–18624. Curran Associates Inc, 2020. https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Duong, L.T., Nguyen, P.T., Di Sipio, C., & Di Ruscio, D., (2020). Automated fruit recognition using efficientnet and mixnet. Computers and Electronics in Agriculture, 171: 105326, 2020. ISSN 0168–1699. doi: 10.1016/j.compag.2020.105326. http://www.sciencedirect.com/science/article/pii/S0168169919319787.

Fawcett, T. (2006). An introduction to roc analysis. Pattern Recogn. Lett., 27 (8): 861–874, June 2006. ISSN 0167–8655. DOI: 10.1016/j.patrec.2005.10.010. doi: 10.1016/j.patrec.2005.10.010.

Han, G., Liu, X., Zhang, H., Zheng, G., Soomro, N.Q., Wang, M., & Liu, W. (2019). Hybrid resampling and multi-feature fusion for automatic recognition of cavity imaging sign in lung ct. Future Generation Computer Systems, 99: 558–570, 2019. ISSN 0167–739X. doi: 10.1016/j.future.2019.05.009. https://www.sciencedirect.com/science/article/pii/S0167739X19306806.

Harris, M., Qi, A., Jeagal, L., Torabi, N., Menzies, D., Korobitsyn, A., Pai, M., Nathavitharana, R., & Ahmad Khan, F. (2019). A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PLOS ONE, 14*, 09. https://doi.org/10.1371/journal.pone.0221339. e0221339.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Heidari, M., Mirniaharikandehei, S., Khuzani, A.Z., Danala, G., Qiu, Y., & Zheng, B. (2020). Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. International Journal of Medical Informatics, 144: 104284, 2020. ISSN 1386–5056. doi: 10.1016/j.ijmedinf.2020.104284. http://www.sciencedirect.com/science/article/pii/S138650562030959X.

Huang, Z., Pan, Z., & Lei, B. (2017). Transfer Learning with Deep Convolutional Neural Network for SAR Target Classification with Limited Labeled Data. Remote Sensing, 9 (9), 2017. ISSN 2072–4292. DOI: 10.3390/rs9090907.

Hwang, S., Kim, H.-E., & Kim, H.-J. (2016). A novel approach for tuberculosis screening based on deep convolutional neural networks. In G.D. Tourassi and S.G.A. III, editors, Medical Imaging 2016: Computer-Aided Diagnosis, volume 9785, pages 750–757. International Society for Optics and Photonics, SPIE, 2016. DOI: 10.1117/12.2216198. doi: 10.1117/12.2216198.

Iovino, L., Nguyen, P. T., Salle, A. D., Gallo, F., & Flammini, M. (2021). Unavailable transit feed specification: Making it available with recurrent neural networks. *IEEE Transactions on Intelligent Transportation Systems, 22*(4), 2111–2122. https://doi.org/10.1109/TITS.2021.3053373

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., & Thoma, G. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging Medicine and Surgery, 4*(6), 475.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S., Thoma, G., Wang, Y.-X., Lu, P.-X., & McDonald, C. J. (2014). Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging, 33*(2), 233–245. https://doi.org/10.1109/TMI.2013.2284099

Jain, A., Mishra, A., Shukla, A., & Tiwari, R. (2019). A novel genetically optimized convolutional neural network for traffic sign recognition: A new benchmark on belgium and chinese traffic sign datasets. *Neural Processing Letters, 50*(3), 3019–3043. https://doi.org/10.1007/s11063-019-09991-x

Jia, L., Alizadeh, R., Hao, J., Wang, G., Allen, J.K., & Mistree, F. (2020). A rule-based method for automated surrogate model selection. Advanced Engineering Informatics, 45: 101123, 2020. ISSN 1474–0346. doi: 10.1016/j.aei.2020.101123. https://www.sciencedirect.com/science/article/pii/S1474034620300926.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2 (4): 230–243, 2017. ISSN 2059–8688. DOI: 10.1136/svn-2017-000101.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12* (p. 1097). Red Hook, NY, USA: Curran Associates Inc.

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology, 284*(2), 574–582. https://doi.org/10.1148/radiol.2017162326. PMID: pmid:28436741genbank28436741.

Leal-Neto, O., Santos, F., Lee, J., Albuquerque, J., & Souza, W. (2020). Prioritizing covid-19 tests based on participatory surveillance and spatial scanning. International Journal of Medical Informatics, 143: 104263, 2020. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2020.104263. http://www.sciencedirect.com/science/article/pii/S1386505620308534.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., & Xia, J. (2020). Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: Evaluation of the diagnostic accuracy. *Radiology, 296*(2), E65–E71. https://doi.org/10.1148/radiol.2020200905. PMID: pmid:32191588genbank32191588.

L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, page 200905, 2020b.

L. Mansilla, D.H. Milone, and E. Ferrante. Learning deformable registration of medical images with anatomical constraints. Neural Networks, 124: 269–279, 2020. ISSN 0893–6080. doi: 10.1016/j.neunet.2020.01.023. https://www.sciencedirect.com/science/article/pii/S0893608020300253.

H.V. Nguyen, E.W. Tiemersma, H.B. Nguyen, F.G.J. Cobelens, A. Finlay, P. Glaziou, C.H. Dao, V. Mirtskhulava, H.V. Nguyen, H.T.T. Pham, N.T.T. Khieu, P. de Haas, N.H. Do, P.D. Nguyen, C.V. Cung, and N.V. Nguyen. The second national tuberculosis prevalence survey in vietnam. PLOS ONE, 15 (4): 1–15, 04 2020a. DOI: 10.1371/journal.pone.0232142. doi: 10.1371/journal.pone.0232142.

P.T. Nguyen, L. Iovino, M. Flammini, and L.T. Duong. Deep Learning for Automated Recognition of Covid-19 from Chest X-ray Images. medRxiv, 2020b. DOI: 10.1101/2020.08.13.20173997. https://www.medrxiv.org/content/early/2020/08/14/2020.08.13.20173997.

P.T. Nguyen, D. Di Ruscio, A. Pierantonio, J. Di Rocco, and L. Iovino. Convolutional neural networks for enhanced classification mechanisms of metamodels. Journal of Systems and Software, 172: 110860, 2021. ISSN 0164–1212. doi: 10.1016/j.jss.2020.110860. https://www.sciencedirect.com/science/article/pii/S0164121220302508.

F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. Scientific Reports, 9, 2019.

Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ, 7*, Article e6201.

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. B., Ayari, M. A., & Chowdhury, M. E. H. (2020). Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access, 8*, 191586–191601. https://doi.org/10.1109/ACCESS.2020.3031384

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., et al. (2020). Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ Digital Medicine, 3*(1), 1–8.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision, 115 (3): 211–252, Dec. 2015. ISSN 0920–5691. DOI: 10.1007/s11263-015-0816-y.

Samuel, R. D. J., & Kanna, B. R. (2019). Tuberculosis (TB) detection system using deep neural networks. *Neural Computing and Applications, 31*(5), 1533–1545. https://doi.org/10.1007/s00521-018-3564-4

Soltanisehat, L., Alizadeh, R., Hao, H., & Choo, K.-K. R. (2020). Technical, temporal, and spatial research challenges and opportunities in blockchain-based healthcare: A systematic literature review. *IEEE Transactions on Engineering Management*, 1–16. https://doi.org/10.1109/TEM.2020.3013507

A. Souid, N. Sakli, and H. Sakli. Classification and predictions of lung diseases from chest x-rays using mobilenet v2. Applied Sciences, 11 (6), 2021. ISSN 2076–3417. DOI: 10.3390/app11062751. https://www.mdpi.com/2076-3417/11/6/2751.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 843–852). https://doi.org/10.1109/ICCV.2017.97

S. Sutoko, A. Masuda, A. Kandori, H. Sasaguri, T. Saito, T.C. Saido, and T. Funane. Early identification of alzheimer's disease in mouse models: Application of deep neural network algorithm to cognitive behavioral parameters. iScience, 24 (3): 102198,

2021. ISSN 2589–0042. doi: 10.1016/j.isci.2021.102198. https://www.sciencedirect.com/science/article/pii/S2589004221001668.

Szegedy, C., Liu, Wei, Yangqing Jia, P., Sermanet, S., Reed, D., Anguelov, D., Erhan, V. Vanhoucke, & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6105–6114, Long Beach, California, USA, 09–15 Jun 2019. PMLR. http://proceedings.mlr.press/v97/tan19a.html.

M. Toğaçar, B. Ergen, and Z. Cömert. Detection of lung cancer on chest ct images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. Biocybernetics and Biomedical Engineering, 40 (1): 23–39, 2020. ISSN 0208–5216. doi: 10.1016/j.bbe.2019.11.004. https://www.sciencedirect.com/science/article/pii/S0208521619304759.

L. Torrey, T. Walker, J. Shavlik, and R. Maclin. Using advice to transfer knowledge acquired in one reinforcement learning task to another. In Proceedings of the 16th European Conference on Machine Learning, ECML'05, pages 412–424, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-29243-8, 978-3-540-29243-2. URL https://doi.org/10.1007/11564096_40.

H. Touvron, A. Vedaldi, M. Douze, and H. Jégou. Fixing the train-test resolution discrepancy: Fixefficientnet. CoRR, abs/2003.08237, 2020. URL https://arxiv.org/abs/2003.08237.

Vajda, S., Karargyris, A., Jaeger, S., Santosh, K., Candemir, S., Xue, Z., Antani, S., & Thoma, G. (2018). Feature selection for automatic tuberculosis screening in frontal chest radiographs. *Journal of Medical Systems, 42*(8), 146.

H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, J., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).

Weiss, K., Khoshgoftaar, T., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*, 12. https://doi.org/10.1186/s40537-016-0043-6

Whiting, P., Rutjes, A., Westwood, M., Mallett, S., Deeks, J., Reitsma, J., Leeflang, M., Sterne, J., & Bossuyt, P. (2011). Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*, 529–536.

C. Xie, M. Tan, B. Gong, J. Wang, A. Yuille, and Q.V. Le. Adversarial Examples Improve Image Recognition. arXiv e-prints, art. arXiv:1911.09665, Nov. 2019.

Q. Xie, M.-T. Luong, E. Hovy, and Q.V. Le. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

Zeng, N., Wang, Z., Zhang, H., Kim, K.-E., Li, Y., & Liu, X. (2019). An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips. *IEEE Transactions on Nanotechnology, 18*, 819–829. https://doi.org/10.1109/TNANO.2019.2932271

N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, and X. Liu. Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip. Neurocomputing, 425: 173–180, 2021. ISSN 0925–2312. doi: 10.1016/j.neucom.2020.04.001. https://www.sciencedirect.com/science/article/pii/S0925231220305385.