Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 20, Iss. 2, pp. 89-103, 2022

MACHINE LEARNING BASED SAMPLING OF X-RAY IMAGES FOR A COMPUTER-AIDED DETECTION OF TUBERCULOSIS *

Fernando Ferreira
Philipp Gaspar
Rodrigo Torres
Carlos Eduardo Covas
Lukas Müller de Oliveira
Micael Veríssimo de Araújo
José Manoel de Seixas
Federal University of Rio de Janeiro
Signal Processing Laboratory COPPE/POLI
emails: {fferreira, philipp.gaspar, torres, kaducovas, lukasmullerdeoliveira, micael.verissimo, seixas}@lps.ufrj.br

Mayara Bastos
Research Institute of the McGill University Health Centre
email: mayara.bastos@yahoo.com

Anete Trajman
Federal University of Rio de Janeiro
Faculty of Medicine
email: atrajman@gmail.com

Abstract – Computer-Aided Detection software relies on annotated data set of X-rays to be developed. The annotation task is time-consuming and requires extensive know-how. This work presents a sampling method to select the most relevant images, which will be annotated for the development of a tuberculosis (TB) screening platform based on machine learning algorithms. The sampling task optimizes the annotation process by reducing the number of images to be analyzed without compromising the diversity and the significance power of the images in the dataset. We developed an algorithm to select images in a dataset to be annotated, based on similarity and dissimilarity measurements of images. Public TB image dataset was utilized to conduct this research. The experiment consisted of a deep learning feature engineering step, followed by topological analysis based on Self-Organizing Map and K-Means. The effectiveness of the process is evaluated at each of its stages: Classification, clustering and the final sampling algorithm which is based on similarity and dissimilarity features.

Keywords – Deep Learning, CNN, SOM, Clustering, CAD

1 Introduction

Tuberculosis (TB) remains a major global health problem, as it is one of the leading causes of mortality worldwide, especially in low-income countries [1]. Before the COVID-19 pandemic, TB was the most infectious disease killer globally. In 2020, there were 5.8 million new TB cases and 1.4 million deaths [2]. The pandemic period has caused serious disruptions in the management of TB, setting back the fight of the disease pace by several years [3].

Medical imaging has an important role in the diagnosis and management of pulmonary tuberculosis (PTB) [4–6], especially in the early stages of the disease, as chest radiography (also known as chest X-ray or simply CXR) is a sensitive screening tool for detecting TB in children and adults who are at higher risk of contamination [2], being recommended by World Health Organization (WHO) as an effective screening for PTB and also as a diagnostic aid to complement bacteriological tests [7].

Nevertheless, the radiographic presentation of pulmonary TB may exhibit diverse and heterogeneous characteristics, making the diagnosis a challenging task. For example, it is important to distinguish between active and inactive TB. It is only possible to achieve the goals of sustainable development goals if we treat both [8].

Although both have distinct markers, there is considerable overlap between their appearances [9]. In addition, the limited access to high-quality digital CXR imaging in low resource sites and a lack of trained readers make the task of image interpretation harder [10].

*This work was funded by CNPQ (process 440129/2020-6) The authors would like to thank CAPES and FAPERJ for their support during this work. Also, the authors would like the acknowledge the researches of the project "Diagnóstico auxiliado por computador para exclusão de tuberculose ativa em contatos de pacientes com tuberculose pulmonar – quebrando a cadeia de transmissão" for all the insightful comments and development suggestions.

CAD software packages have been used to support radiologists during clinical practice in detecting potential abnormalities on diagnostic exams [11]. In some applications, the CAD software is also used to indicate the likelihood that a CXR image represents a specific disease process [12]. They often rely on Artificial Intelligence (AI) algorithms to perform the image analysis. AI usage to detect TB in CXR has progressed significantly in the past 30 years [12–17].

Recently (2021), WHO has recommended that CAD applications could be used as a complementary alternative to human interpretation of digital CXR for Tuberculosis screening and triage of adults aged 15 years or more [2].

Annotated databases of images are fundamental resources for building CAD applications based on AI algorithms as the model training task depends on them [18, 19]. Therefore, the quality of the annotated dataset has a direct impact on the quality of the CAD software. Since creating a large, annotated medical image dataset is not easy, most researchers rely on the public available CXR datasets [20]. However, in most cases, the X-ray imaging annotation on those datasets is not accurate and/or certificated [21].

This imposes a challenge that precedes the development of the models used in CAD: in order to use qualified annotated data for the model training, a reviewing annotation process must be conducted. This process is time-consuming and expensive as it requires extensive domain knowledge to catalog and index the images manually.

Therefore, the process of sampling images to enter the new annotation phase must follow a selection criterion that prioritizes high diversity images to maximize the data information. In this context, this work presents a sampling method to select images based on similarity and dissimilarity criteria to decrease redundancy and maximize representativeness in the CXR images selected for the annotation process.

In the first part of the analysis, we trained a Convolutional Neural Network (CNN) [22] to perform binary classification of CXR images into TB or not TB. After the training phase, we discarded the last layer of the network to get its embeddings, found during the training process, of the second to the last layer. Then, the embeddings were projected into a Self-Organizing Map (SOM) [23] responsible for transforming those samples into a 2D grip map. The final part of the analysis consists of a clustering algorithm used for sampling CXR images based on similarity and dissimilarity criteria: each image is assigned to a SOM unit and these units are clustered according to their proximity to their neighbors. In such a manner, it is possible to collect different kinds of images from the map regions. Those who seem to be "misplaced" (neighbors of different diagnoses) are considered dissimilar. Otherwise, when most of the neighbors are from the same class, the CXR may be considered good representatives of their clusters and will be marked to be of a similar type. In the end, the process will create a list with 40 similar images and 60 dissimilar images.

This work is part of an ongoing project. Researchers are expected to access other public datasets shortly and new annotation efforts might be required. For this purpose, an experiment to assess the model generalization capability is presented and discussions are made in the results section.

2 Background Material

This section details the proposed signal processing pipeline, which, as shown in Figure 1, consists in the following 4-steps pipeline:

- convert CXR images to grayscale and resize them to a fixed size.
- extract a set of discriminative features in order to exploit different aspects of x-ray images.
- project data in a non-linear way for both reducing dimensionality and representing the data set in fewer elements, preserving the topology of the observations.
- Group observed samples (or projections) into clusters according to similarity measures.

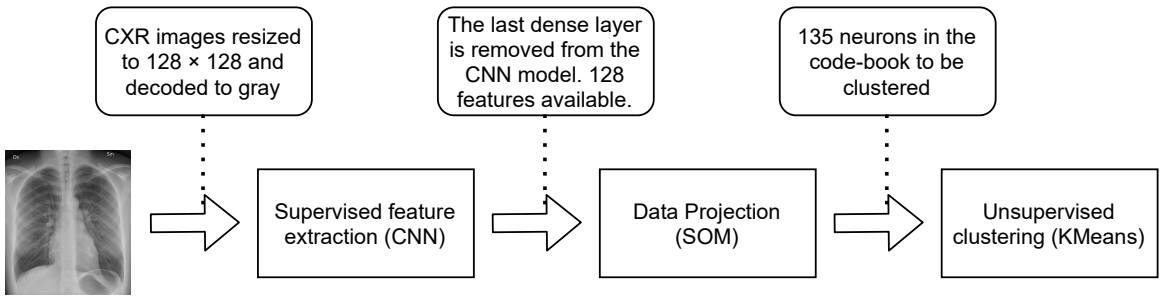


Figure 1: Proposed sampling method pipeline.

2.1 Feature Extraction

Feature extraction is one of the most important steps of medical image processing, which requires extensive domain knowledge [24]. For many unsupervised learning techniques, replacing raw images with features extracted by convolutional neural networks (CNN) leads to better results as this kind of architecture can efficiently produce a set of discriminative features without any expert guidance [24, 25].

2.1.1 Convolutional Neural Networks (CNN)

CNN is a class of deep neural network architecture widely used in image analysis and it consists of an input layer, hidden layers, and an output layer. This specialized type of neural network performs convolution operations on images due to the way weights are shared throughout the network. After passing through a convolutional layer, the image becomes abstracted to a feature map, also called an activation map. At each layer, the input image is convolved with a set of K kernel $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_k\}$ and added biases $\mathcal{B} = b_1, \ldots, b_2$, each generating a new feature map $\mathcal{X}_{\|}$. These features are subject to an element-wise non-linear transform $\sigma(.)$ [26]. This deep learning architecture does not need to learn separate detectors for the same object occurring at different locations of an image. It is capable of learning the parameters and extracting the global and local features that are more discriminative in an image [27]. For these reasons, CNN provides the flexibility of extracting intrinsic and discriminative features from X-ray images [24] and thus, became a popular design approach deployed in several recent works about tuberculosis automated detection [15, 28–31].

2.2 Data projection methods

Projection methods are mainly meant for reducing the data dimensionality by representing the observation in a subspace (lower number of directions) that concisely describes the data structure and features are preserved faithfully [32]. Principal Component Analysis (PCA) [33] is widely used due to its simplicity and the possibility to reduce the data dimensionality in a controlled way: each extracted component (dimension) is ranked by how much data variance it represents.

In summary, PCA searches in an unsupervised way for uncorrelated directions. The eigenvectors of the data covariance matrix represent the principal directions for which the variance of data has the maximum values. The corresponding eigenvalues define how much energy the component retains. In Equation 1, if the principal directions are represented by u_j, we can define a linear orthogonal transformation of data x as:

$$p_j = u_j^T x = x^T u_j, j = 0, \ldots, N-1 \tag{1}$$

where p_j is the data projection onto a principal component. In order to reduce the original data dimension, one may use only the major n projections, discarding the projections of smaller variance. It is important to notice that PCA only considers second-order statistics. Here, the PCA is combined with the Self-Organizing Map (SOM), a computational data analysis method that produces nonlinear mappings of data to lower dimensions [34].

2.2.1 Self-Organizing Map

The SOM algorithm consists of an unsupervised trained neural network with the Kohonen layer [23]. The algorithm uses a similarity measure (usually based on the Euclidean distance) between data samples to embed a low dimensional space (typically two dimensions) into the original data. The code-book maps the data space into a 2D-grid. The result is a topological organization of the input data where their most relevant aspects are preserved, revealing hidden structures throughout the non-linear mapping.

The SOM usually computes the Euclidean distance of the input vector to each neuron, and find the winning neuron, denoted neuron, using the nearest-neighbor rule (competition phase). The winning node is called the excitation center and it determines a neighborhood of excited nodes (cooperation phase). All the input vectors that are within the neighborhood of such winning neuron adjust the weights in order to strengthen its response (adaptation phase) [35].

The input data space is fully connected to each neuron from the Kohonen layer, and the weights are computed iteratively using:

$$w_j(n+1) = w_j(n) + \eta(n) h_{ij}(n)(x(n) - w_j(n)) \tag{2}$$

where $\eta(n)$ stands for the learning rate and $h_{ij}(n)$ is the neighborhood kernel. In this application, we used a Gaussian kernel:

$$h_{ij}(n) = \exp\left(\frac{-d_{ij}^2}{2\sigma^2(n)}\right) \tag{3}$$

where d_{ij}^2 denotes the similarity measure, and $\sigma(n)$ is the monotonically decreasing width of the kernel. In the end, after the training procedure, the network outputs are calculated for each neuron using Equation 4.

$$u_i = x^T w_i \tag{4}$$

where the vector w_i represents the weight that connects the input data to the neuron i. Hence, the vectors $w_{i,i=1,m}$ form the SOM code-book.

The obtained map offers powerful tools for data exploration due to its bi-dimensional (in most cases) nature, which allows identifying cluster borders, projection directions, and possible dependencies between variables [32].

2.3 Data clustering

Clustering methods are widely used for grouping observations into smaller subsets (clusters) defined by a similarity measure. In other words, samples belonging to a cluster should be as much as similar to each other and quite different from observations from other clusters.

The most common clustering technique is the k-Means algorithm, which is described in [36]. The SOM can be viewed as a clustering algorithm that produces a set of clusters organized on a regular grid [35].

2.4 Clustering consistency measures

The k-Means clustering procedure does not provide any measure of how consistent the estimated clusters are. Hence, the silhouette index [37] provides a measurement of the consistency of the estimated clusters through the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

where $a(i)$ is the average dissimilarity between the $i-th$ sample from a given cluster and the other samples in the same cluster and $b(i)$ is the lowest average dissimilarity of the $i-th$ sample to the remaining clusters. The above expression can be rewritten as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ a(i)/b(i) - 1, & \text{if a(i) ¿ b(i)} \end{cases} \tag{6}$$

It can be seen that $-1 < s(i) < 1$. If the $i-th$ sample is well-represented in its cluster, then $s(i)$ is closer to 1. Otherwise, $s(i)$ is closer to -1. Data samples with silhouette values close to zero are in the border of two or more clusters. The average silhouette index, considering all data samples, is used to measure the performance of the clustering configuration.

2.5 Neighborhood similarity

An important aspect of SOM is the neighborhood similarity. This is active by measuring the distance among nearby neurons. In our case, we are also interested to measure the similarity related to the diagnosis status (TB or not TB) between the input vector and other vectors that activate the same neuron and its neighbors. Therefore, we need to define a neighborhood function.

First, the number of data points of a given class is computed for each neuron. This way, it is possible to understand the empirical probability of the input vector to be labeled as TB and not TB. Then, the Gaussian function is used to propagate the influence of the input vector to the neighbors. Thus, the neighborhood similarity function is defined as:

$$S^A = \sum_{i}^{N} \sum_{j}^{M} \mathcal{C}_{ij} \times \mathcal{N}(i, j, \sigma) \tag{7}$$

where i and j are coordinates of the SOM, $\mathcal{C}_{ij}^A$ is the number of data points of class A that activate the neuron (i, j), $\mathcal{N}(i, j, \sigma)$ is the Gaussian function centered in coordinate (i, j), and S^A is the neighborhood similarity matrix computed for class A.

2.6 Quality measures for comparing SOMs

When training multiple SOMs, comparing them is a difficult task. Usually, only some topological differences are assessed via visual inspection or a couple of measures are computed to compare vector quantization. Both ways are not enough for studying map stability across a range of training approaches or to further analyze data. Quality measures based on data co-location and shift analysis of output space mapping can be used for an analytical way of performing the systematical comparison among different SOMs [38].

Data Shifts Analysis (DSA) This method can be used to find out how stable the mapping is, and how steadily a data vector is put into the neighborhood of other vectors on different SOMs [38]. In other words, it measures how much of the data topology on the map is caused by the attributes of data, and how much are due parameters or initialization.

Cluster Shifts Analysis (CSA) This method is similar to the Data Shifts Analysis, but compares SOMs on a more aggregated level, by comparing clusters in the SOM instead of single units or neighborhoods [38].

In this work, several SOM maps were trained, especially for studying their generalization capabilities. However, as our method consisted in using different subsets of the data in each training (through cross-validation procedure), it is not possible to perform the DSA. The CSA is more suitable for our purpose, as K-Means clustering has already been used for clustering the SOM units on a more aggregate level. Moreover, as every model was created using overlapped data from other models, it is expected that clusters present a more stable behavior [39].

The CSA is computed by extracting the same number of the cluster on two SOMs. The clusters found in both SOMs are linked to each other. The highest matching number of data points for pairs of clusters on both maps, the higher is the confidence that the two clusters correspond to each other [38]. Here, the Jaccard similarity coefficient [40] is used to calculate the confidence level.

3 Experiment Setup

This section describes the proposed pipeline to deliver the most representative CXR images from an extensive database will be used in a later annotation task. This pipeline is designed to select images based on similarity and dissimilarity measurements.

Firstly, the CXR images are resized to 128×128 and decoded to a gray-scale single channel, where each value varies between 0.0 (black) and 1.0 (white). The resulting vectors are normalized by the standard deviation. Then, deep learning-based feature engineering is performed using a CNN architecture.

The CNN setup uses a 6-layer network structure, which is consist of four convolutional layers and two fully-connected layers, as shown by the schematic diagram in Figure 2. Dropout and ReLU are deployed to address over-fitting and convergence issues [41]. The model was trained using Adam optimization algorithm [42] with 10 epochs and batch size equals to 64. The optimization parameters were learning rate equals to 0.001, beta1 as 0.9 and beta2 as 0.999.

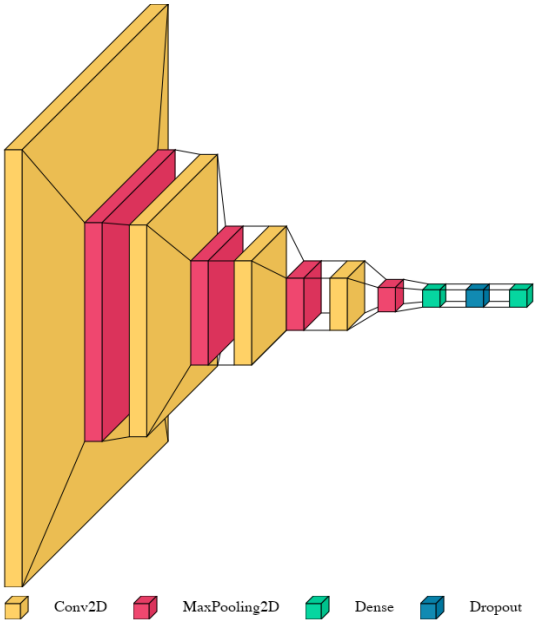


Figure 2: The proposed CNN architecture proposed has 6 layers: [1] Conv2D (32 filters), kernel (3×3), ReLU, Max Pooling (2×2) [2] Conv2D (64 filters), kernel (3×3), ReLU, Max Pooling (2×2) [3] Conv2D (128 filters), kernel (3×3), ReLU, Padding Same, Max Pooling (2×2) [4] Conv2D (128 filters), kernel (3×3), ReLU, Padding Same, Max Pooling (2×2) [5] Fully-connected (128 units), ReLU, DropOut (0.5) [6] Fully-connected (1 unit), Sigmoid

The available dataset was divided into 10 separated partitions using a stratified cross-validation split [43] procedure assuring that we have well-balanced classes in each train and validation subset. Also, an early stopping strategy was adopted through motoring the validation set using the SP (Equation 8) index [44] at each iteration.

$$SP = \sqrt{\sqrt{P \times (1-F)} \times \left(\frac{P + (1-F)}{2} \right)} \tag{8}$$

where P denotes the true positive rate and F the false positive probability.

After the training process, the last dense layer is removed from the CNN model and the CXR images are fed forward through the network. The efficiency of the network architecture is calculated by computing the mean SP vaie of the validation set for each partition model. In order to select one single model por operational use, the model with the highest mean SP for the entire dataset is used. The representations found by this procedure will have size $n_{\text{samples} \times 128}$, since the dense layer has 128 neurons and will be used in subsequent models: PCA and SOM. At this point, those models were trained also using the entire dataset, with no generalization strategy. A more in-depth analysis of this choice will be addressed after the operation model building phase.

PCA is used to reduce the dimensionality of the embeddings found by the CNN. The SVD decomposition is computed using the divide-and-conquer approach for computing the SVD [45]. The first 10 components were enough to capture approximately 100 % of the variance of the embeddings and thus used for analysis.

The SOM model built for exploring the nonlinear statistical relationships among the extracted features were optimized to a 15×9 (135 neurons) grid using a rectangular topology. The number of neurons is calculated from the number of data points of the training dataset using the Equation 9, as proposed by [46]:

$$M \approx \sqrt{5}I \tag{9}$$

where M represents the number of neurons, rounded to the nearest integer and I is the number of CXR images.

The SOM network weights were initialized by spanning the first two principal components. This initialization does not depend on random processes, making the training process much faster [46]. As training parameters for building the map, the learning rate was set to 0.5, the spread of neighborhood function was equal to 2 and the maximum number of 1000 epochs. Euclidean distance was used as an activation function and the Gaussian function was used to weigh the neighborhood of a position in the map.

Later, the SOM codebook was clustered using the K-Means algorithm varying the number of centroids from 2 to 20 centroids and the results were compared using the silhouette index. At this point, we are not interested in finding the very best clustering model (with higher silhouette values).

The clustering configurations are exploited to understand the intrinsic structures evidenced by performed non-linear mapping. This task aims to select samples according to similarity and dissimilarity parameters and the silhouette index will help to understand how well-fitted the given model is and how well the clustered samples are conformed inside it. In other words, how similar a given sample is to its neighbors.

The next section will show the obtained analytic results.

4 Dataset

Some of the difficulties in building CAD software for lung diseases are due to the small datasets (¡10^3) publicly available, which also presents some additional artifacts created by radiologists, like text and symbols written in different parts of the image. Nevertheless, this work uses the Shenzhen dataset collected at te Shenzhen No.3 People's Hospital [47], containing normal and abnormal chest X-rays with manifestations of TB, also including associated radiologist readings and initial classification labels used to train the CNN.

The used data set contains 662 frontal chest X-rays, of which 326 were labeled as normal cases and 336 were cases with manifestations of TB, including pediatric X-rays (anteroposterior). The chest X-rays are from outpatient clinics and were collected from the Shenzhen No.3 Peopleś Hospital, Shenzhen, China. The images are provided in PNG format as 12-bit gray-level images and their size is approximately 3000×3000 pixels.

5 Results and Discussion

To obtain the performance measurements of the CNN, the entire dataset (train and validation) is passed throughout the models trained in each phase of the cross-validation procedure.

Table 1 shows the CXR image classification performance (average value, error bars from one RMS value) considering TB or non-TB detection using the test partitions from the cross-validation procedure. Here, sensitivity refers to the efficiency for the correct classification of TB images, while specificity stands for the efficiency in terms of the not TB identification.

Envisaging the practical operation from a single model, the selection was made through evaluating the SP index when models were fed from all the datasets. Next, results are from such a selected operation model.

Table 1: CXR image classification performance for TB or non-TB detection using the test partitions from the cross-validation procedure.

SENSITIVITY	78.2% +- 10.0%
SPECIFICITY	88.0% +- 5.2%
SP	84.4% +- 3.9%

The high standard deviation, shown in Table 1 is due to the fact that the dataset used for developing the models was not assisted by any data quality modeling. Thus, there are images with artifacts, decentralized lungs, lateral images, CXR of children, among others. This might account for fluctuations and lower efficiency values. When these images fall, within the cross-validation draw, in the test set, the network may find it difficult classifying them. Besides, such data quality issues increase statistical fluctuations along model training.

All results shown in Table 1 were obtained using the threshold found by the SP index, which aims at a balance between the false positive and true positive rates. Better performance in terms of sensitivity could be achieved by choosing a different threshold. This is illustrated in the Figure 3a, which shows the ROC curve averaged across all 10 folds used in the cross-validation process. Anyhow, as the overall aim was to sample images in the dataset when diversity and similarity are of concern, a balanced figure between sensitivity and specificity would provide the best threshold, as is the case when we apply the SP index as the figure of merit. Thus, the embeddings provided by CNN would reflect such a balance between TB and non-TB cases.

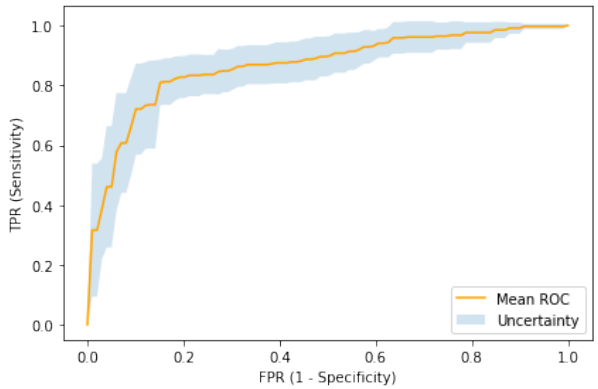
A Transfer Learning technique was also applied, using the same cross-validation partitions as the CNN trained from scratch. A ResNet [48], with loaded ImageNet weights, was retrained for the CXR images with all its layers frozen except the last two. The results are sumarized in the Table 2 and the ROC curve is shown in Figure 3b.

The classification performance of the Tranfer Learning approach is similar, within the uncertainty bands, with the CNN trained from scratch. However, the medical field may find it suspicious to use images from other areas to classify diseases based on images of patients. Hence, we decided to use the embeddings of the network trained only with radiographic images.

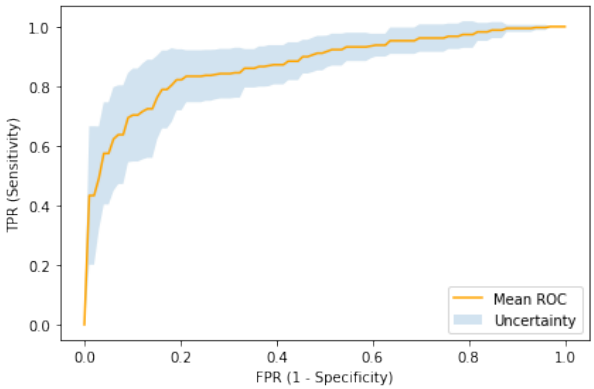
The two first principal components were computed and together they represented 80 % of the PCA explained variance and the ratio between them is approximately 5.39. As said before, the two components are used for initial determination of the SOM topology (15×9) neurons).

Table 2: CXR image classification performance for TB or non-TB detection for the ResNet training using the test partitions from the cross-validation procedure.

SENSITIVITY	78.0% +- 9.3%
SPECIFICITY	88.9% +- 6.5%
SP	84.8% +- 5.5%



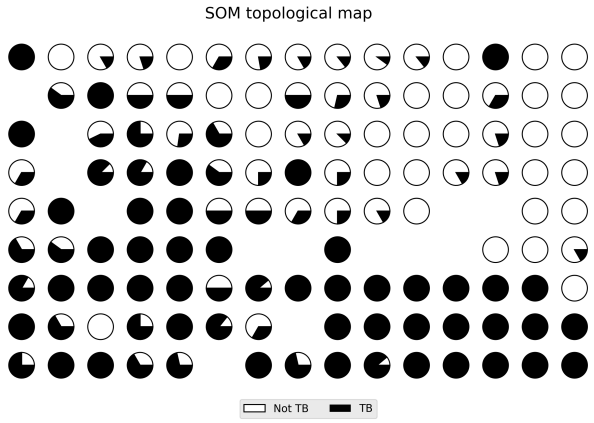
(a) ROC Curve averaged across the cross validation folds for the trained CNN



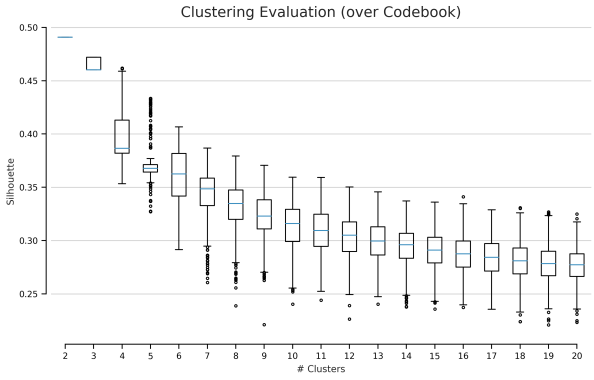
(b) ROC Curve averaged across the cross validation folds for the trained ResNet.

Figure 3: ROC Curve averaged across the cross validation folds for the trained CNN and ResNet.

Then, the features are projected over the SOM map and the result can be seen as the graphical representation displayed in Figure 4a. Each neuron is represented by a circle (neurons with no hits are hidden from the graphic) which shows the ratio between CXR images of patients with and without TB that hit the neuron. Filled circles represent neurons hit by vectors extracted from the CXR of TB patients. On the other hand, hollow circles demonstrate neurons reached only by vectors from CXR images with no TB detected.



(a) Graphical representation of the SOM topological map. Each neuron is represented by a circle and colored according to the ratio of "not TB" and "TB" hits.



(b) K-Means clustering evaluation over the SOM code-book. Each clustering setup was initialized 100 times. In the box plots, the silhouette index variance is shown for each clustering model.

Figure 4: Results of SOM projection and K-Means clustering over the SOM code-book.

The graph clearly shows a cross-sectional band (from upper left to the lower right) which divides two regions, one with mostly CXR images of TB-infected patients and the other with non-TB patients. Yet, both regions presented neurons hit by both kinds of patients. These "mixed neurons" might represent images with a high degree of dissimilarity, as they were excited by images from distinct classes; while neurons that hit only one type of patient might represent a degree of similarity, especially when located in regions with similar neighbors.

At this point, K-Means is applied in order to surface the intrinsic relations. Figure 4b shows the silhouette index values for several clustering initializations. The number of centroids was varied from 2 until 20 and each setup was initialized 100 times. The best-fitted models have 2 centroids (as could be foreseen from the cross-sectional band analysis described earlier) with silhouette index slighted over 0.50, with minimal variance. Models with 3 and 4 centroids present a silhouette over 0.42 for all initialization,

with picks of 0.48 and 0.47, respectively. These three configurations were considered for the exploitation analysis since from 5 clusters the silhouette value drops to below 0.40 on average, with little variation as we increase the number of centroids.

5.1 CXRs selection procedure

At this stage, it is time to exploit the intrinsic relations uncovered by all nonlinear methods applied in sequence. For this purpose, the following characteristics will be observed: the samples best conformed within a cluster (indicating similarity); samples with higher silhouette values, but placed in a cluster with neighbors labeled differently (indicative of dissimilarity); and samples located at the borders of the clusters, in transition areas with little activated neurons (indicative of dissimilarity). This analysis is performed on the best models with 2, 3, and 4 centroids. In this way, samples were chosen until completing 50 samples.

Figure 5 shows the SOM Maps for different clustering settings and each one allowed the detailed analysis according to different aspects, describe as follows:

A) Figure 5a displays 2 clusters and TB patients are pointed by hollow circles varying their sizes according to the silhouette value. The bigger the circle, the better the CXR is fitted inside the cluster. The predominance of circles in the cluster on the bottom of the map is visible, mostly with values between 0.4 and 0.7 (except objects located in the border area between the clusters). This region is mostly composed of neurons that were excited by images of patients diagnosed with tuberculosis. Therefore, objects (25 images) with higher silhouette values with this class can be chosen as good representatives for the criterion of similarity with TB. Figure 8a displays an example of this group of images.

B) Figure 5b also displays 2 clusters, but it shows only non-TB patients (cross marker sized proportional to the sample's silhouette values). As expected, the crosses are mostly located on the top side of the map. There are some objects located at the Cluster 0 with high silhouette values. These are interesting cases, as it would be expected that only images of TB patients were in the center of this cluster. The clustering result shows some features turn non-TB patients similar to infected ones. Hence, these images (like the ones displayed by Figures 8b and 8e) are good examples of dissimilarity for non-TB images and, for this reason, will be picked for the final sampling. (15 images)

C) Figures 5c and 5d exhibit the SOM map split onto three clusters. It is possible to notice neurons presenting well-fitted samples for both TB and non-TB patients, especially in the center part of the $Cluster1$. Those samples are interesting for representing images that don't share similarities with the target classes. This way, they were picked as dissimilarity examples. (24 images)

D) Figures 5e shows the map divided into four clusters. Cluster 3 shows some samples of TB patients in a region mainly activated by non-TB ones. Therefore, those are sampled as exemplary images with high dissimilarity to other TB images. One example may be seen in Figure 8c. (8 images)

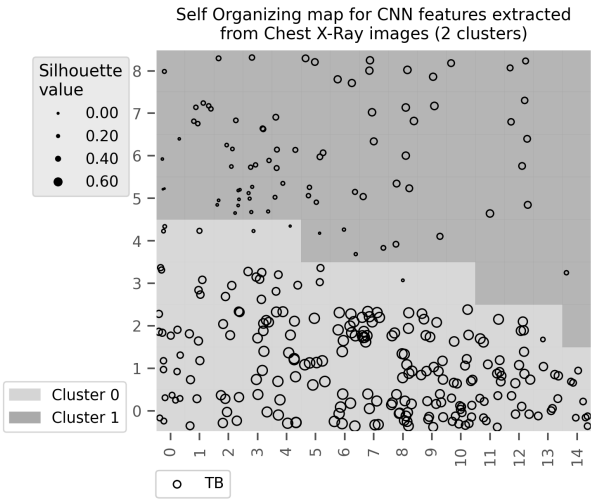
E) Finally, Figure 5f presents the non-TB images split into the same four clusters. Cluster 2 is located in a region where no CXR image labeled as a TB patient was projected. All samples are from non-TB patients, with high silhouette values of around 0.6. These samples might present relevant features to distinguish and differentiate them from positive-TB X-Rays. Hence, they are picked for final sampling by their similarity to non-TB samples. An example may be seen in Figure 8d (16 images).

After the sampling, one hundred CXRs images were chosen. The location of these images on the topological map are shown in Figure 6, which summarizes the results of the clustering analysis. Neurons are colored according to the distance computed to their neighborhood. The four-cluster setup is highlighted by ticker edges and hatched areas. Not sampled images are marked as small dots, making them easier to compare To the visualizations presented in Figure 5.

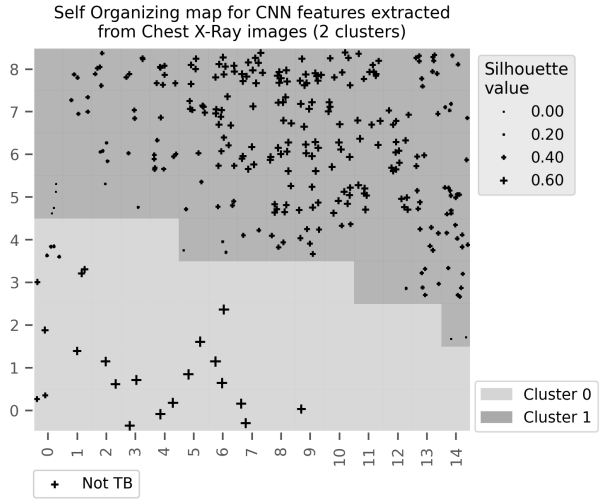
Selected TB patients are represented by red markers and non-TB uses blue markers. It is easy to observe in Figure 6 that the analytical procedure was capable of selecting images from all the clusters, depicting different areas from the map, which should mean a good statistical representation from the dataset, thanks to SOM analysis properties [49]. TB patients sampled due to the similarity criterion are concentrated on the center of the cluster 1, while the ones sampled due to the dissimilarity criterion are spread throughout other clusters. On the other hand, the non-TB patients are located in the left (dissimilarity criterion) and upper right (similarity criterion).

Figure 7 shows the similarity index for each selected image to each class. It is possible to observe that images selected by the similarity criteria (first and second sections of the horizontal axis) present a way higher probability of being from the same class than their neighbors. Another way around, the dissimilarity criteria (sections 3 and 4 of the horizontal axis) picked images with lower similarity to their classes.

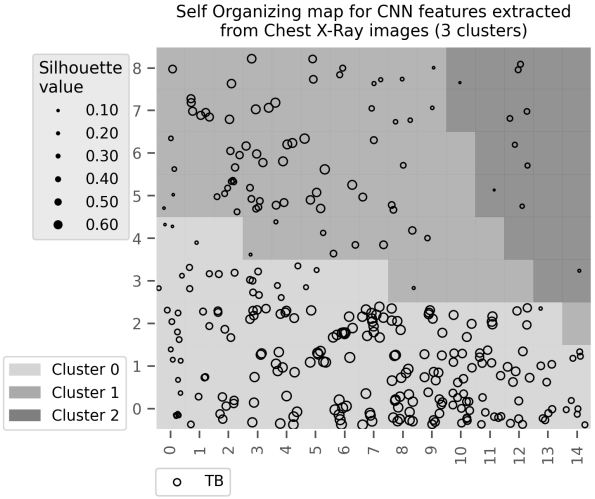
Figure 8 shows sample images selected through the explained analytical exploitation. These CXRs were evaluated by a pair of radiologists in order to establish a validated annotation for the baseline. Some considerations can be made: both agreed that Figure 8a presents a patient with typical features of Tuberculosis; Figure 8b, on the other hand, presents an image that was originally labeled as a non-TB patient, however, the experts understood that it was a case of inactive TB. Something similar occurs with Figure 8e, where the original label indicates TB and there is a consensus that the image depicts a patient also with inactive TB. Figure 8c shows an image that was originally labeled as active TB, but the radiologists couldn't determine the TB activity. And, Figure 8d shows an image that was originally labeled as non-TB and the experts concurred labeling it as a normal non-TB patient.



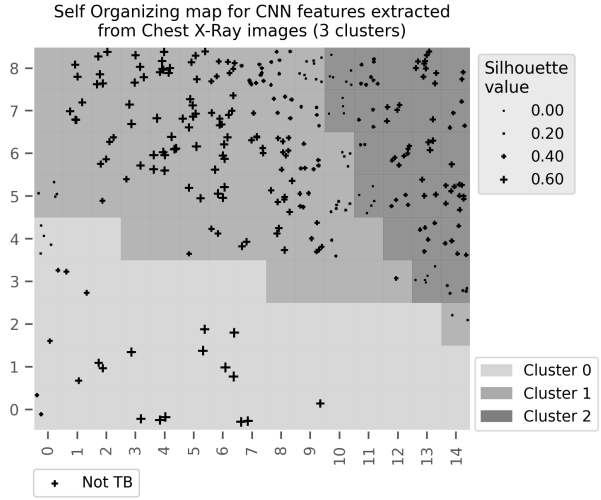
(a) Two clusters - TB patients



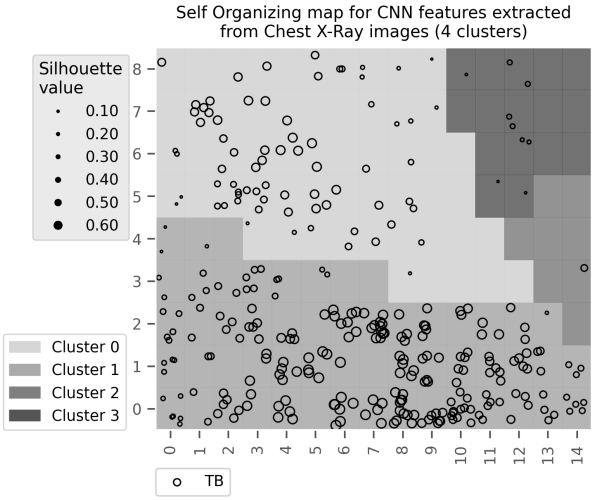
(b) Two clusters - Non-TB patients



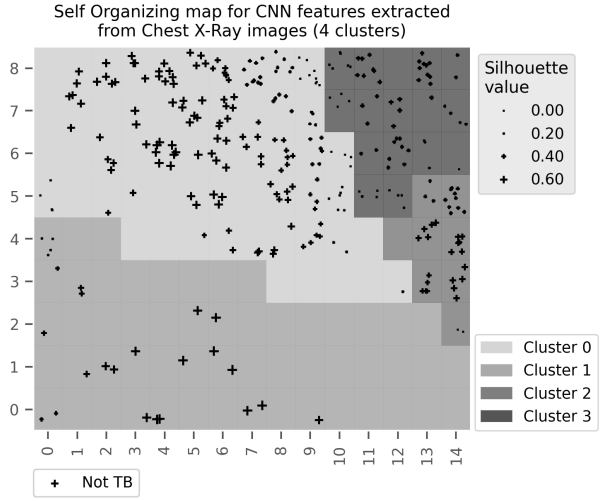
(c) Three clusters - TB patients



(d) Three clusters - Non-TB patients



(e) Four clusters - TB patients



(f) Four clusters - Non-TB patients

Figure 5: Self-Organizing map for the features extracted using the CNN network. The images shows how TB and non-TB patients are projected onto the map and how different clustering models divide the SOM code-book.

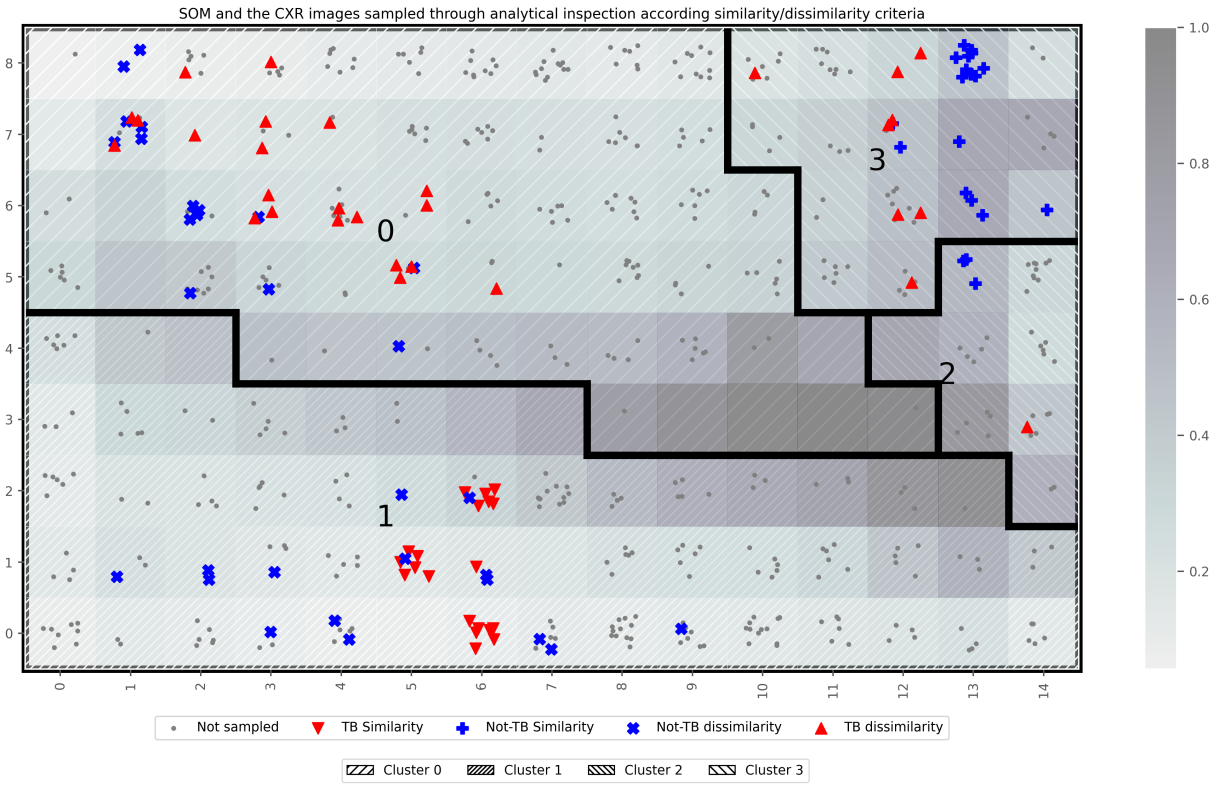


Figure 6: Sampled CXR images through the analytical exploitation according to the similarity/dissimilarity criteria and their location in the topological map. The normalized distances among a neuron and its neighbors are shown by the gray-scale heatmap. The 4 clusters setup is highlighted (edges and hatched area). TB patients are represented by red markers and non-TB uses blue markers. Different markers are used for evincing similarity and dissimilarity.

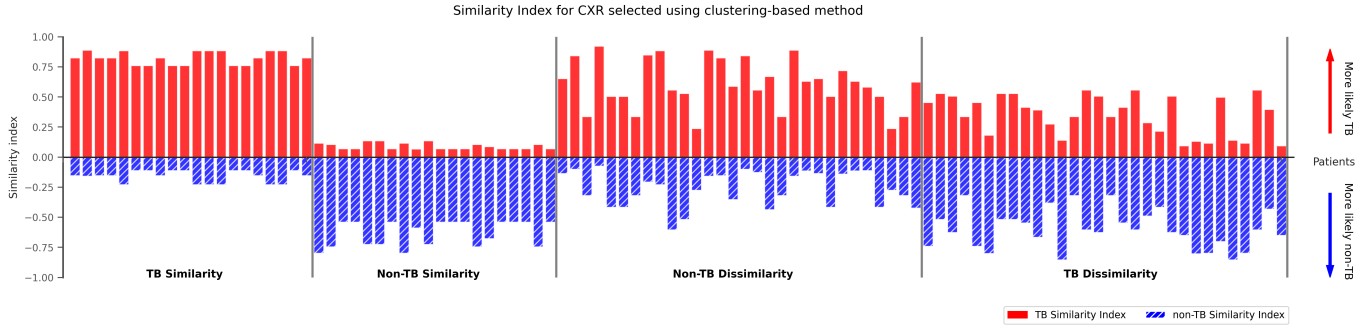
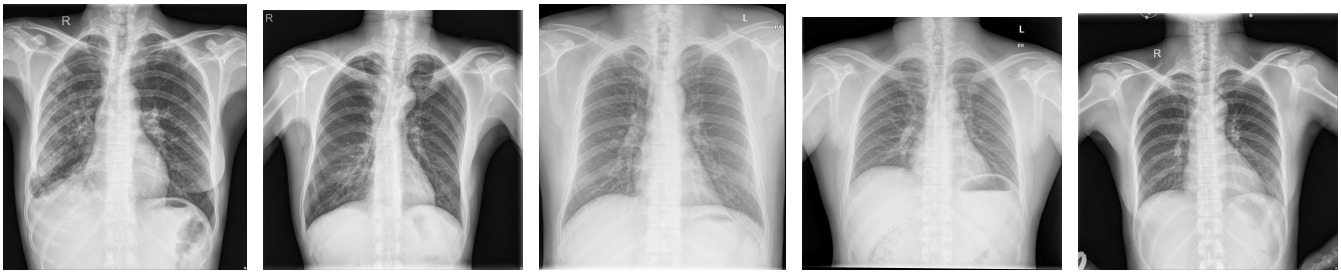


Figure 7: Neighborhood similarity index for the selected CXR images. The x-axis represents the images sorted selection group and y-axis the normalized measure.



(a) Neuron: (5, 1) (b) Neuron: (5, 1) (c) Neuron: (12, 6) (d) Neuron: (3, 6) (e) Neuron: (14, 3)

Figure 8: Examples of Chest X-Ray images selected by similarity and dissimilarity patterns. The coordinates indicates the position in the map shown in Figure 6.

5.2 Model generalization

In the previous section, the unsupervised learning part of the model has used the entire dataset for extracting the clusters. This approach aimed to build an operational model, capable of sampling a list of the most representative images for being delivered to the radiologists.

After the sampling task was completed, one question was raised: how this method would generalize to other datasets? For answering this question, another experiment was set. This time, a different map was created for each step of a 10-fold training procedure. For each step, the 10-folds used for creating the CNN models were used for training the SOM (using 9 folds for building the model and reserving one for testing). Therefore, 10 maps were created. Figure 9, shows the full-data map and one of the 10 maps created, as an example. All points highlighted represent images selected in the previous method. The displayed images are part of the test set for a specific fold configuration (Fold 1).

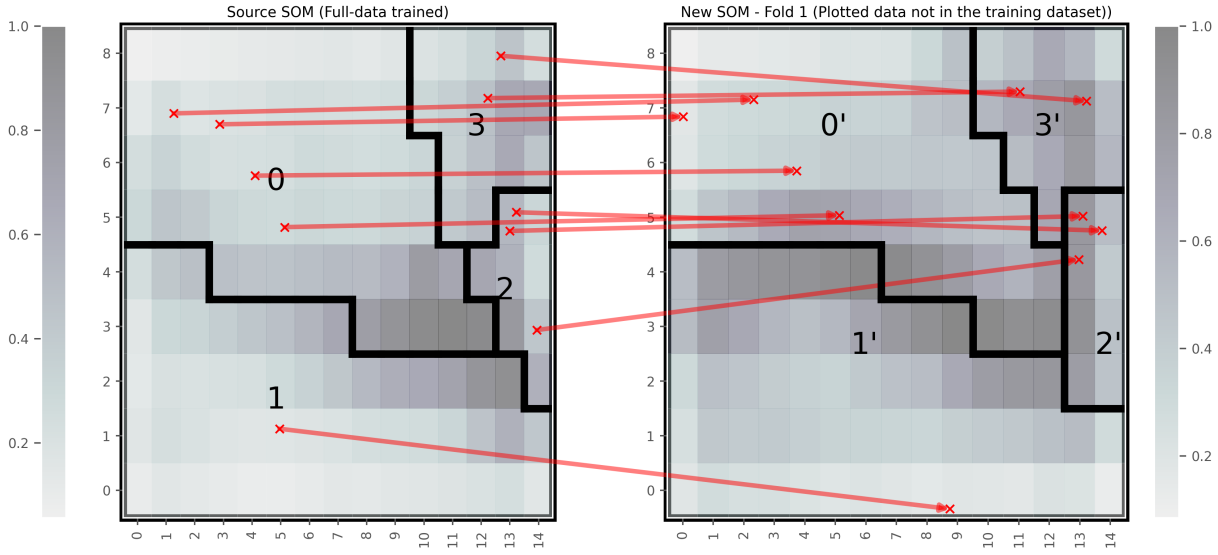


Figure 9: The operational map (full-data trained map) versus one of the ten maps created during the 10-fold cross-validation. It's possible to observe that the map on the right preserved the topological structure of the full-data map.

It is possible to observe that the newly created map preserved the topological structure of the original map. Indeed, the quantization error mean for the 10-fold training was 1.71 ± 0.02 which is similar to the 1.76 obtained in the full-data model. Continuing the visual inspection, it is possible to notice that the data points are distributed similarly to the first map. Comparing the 4-cluster setup for both formations, none of the displayed images have migrated from one cluster to another.

Figure 10 shows the results of the cluster shift analysis, using four clusters as reference. The operational model is used as a source for the analysis and all models created for each validation fold are used as a target. The bar plot shows the mean value for the Jaccard Similarity coefficient between the source model and each target model for the validation dataset. The coefficient equals 1 when all input vectors grouped in the same cluster on the source model are present in the same cluster on the target model. Otherwise, the coefficient equals 0 when none of the images from a cluster source model are represented in its target model corresponding cluster. The confidence interval is represented by the error bars and it is inferred using bootstrapping. The larger clusters (0 and 1) have the highest similarity coefficient. The smaller cluster (2) has the lowest similarity coefficient and the larger confidence interval due to the under-sampling of its images within some of the folds (in the final sampled list there are only five images, considering the entire dataset). For all the clusters, approximately 80% of the images are represented in the corresponding clusters n the source and target models. This result gives an analytical measure of what was perceived through the visual inspection presented in Figure 9: samples are stable across corresponding clusters in different models. This is indicative that the new models preserve the topological structure of the original map.

Moreover, Figure 10 presents a line chart that shows the mean values of the Jaccard Similarity coefficient considering only the sampled images (that were available in the validation dataset). It is important to notice that several folds didn't have any sampled image representing the cluster 2, and these empty folds were not considered for the analysis as the similarity coefficient would be 0.

As a result of the visual inspection analysis combined with CSA, it is possible to affirm that the models built through the cross-validation were able to preserve the topological structure and clustering of their SOM units produced similarly clusters to the ones extracted in the full-data model. The sampled images, whose were originally located at the central part of their clusters in the original map, haven't moved from one cluster to another.

Continuing the analysis, it is necessary to understand whether the data are kept in neighborhoods similar to those found in the operational model. This is a step forward from the previous analysis because by increasing the granularity of the analyzed groups, each individual is characterized by the individuals around it. Even if a sample remains in a corresponding cluster, it may migrate within to a region with intrinsic characteristics different from those found in the original map.

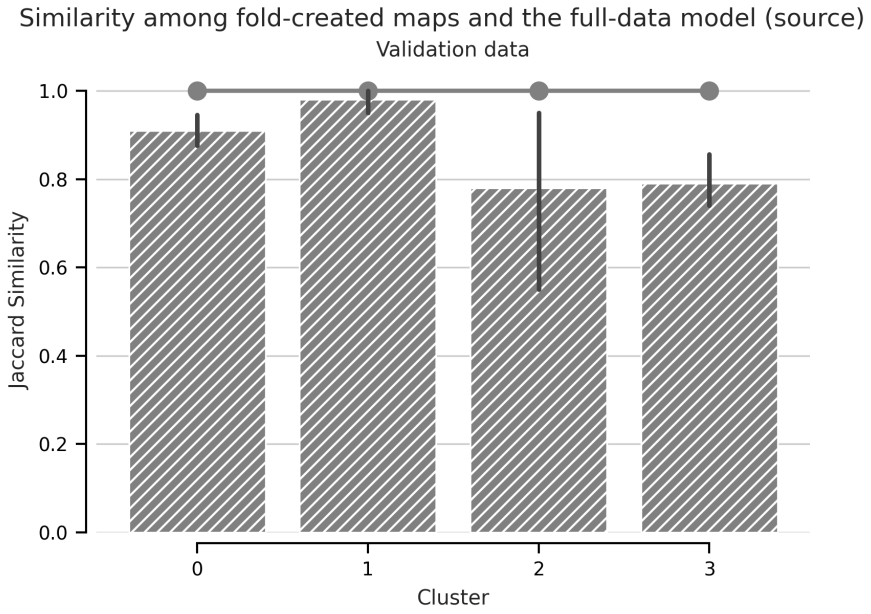


Figure 10: Bar plot represents an estimate for the Jaccard Similarity between the operational model and each model built through the cross-validation process. The line chart represents the similarity coefficient only considering sampled images from the operation.

Figure 11 shows the similarity index for each sampled image to each different selection criteria, considering only the models where those images are on the validation set. It may be observed that the samples are distributed in similar neighborhoods like the ones found in the operational model.

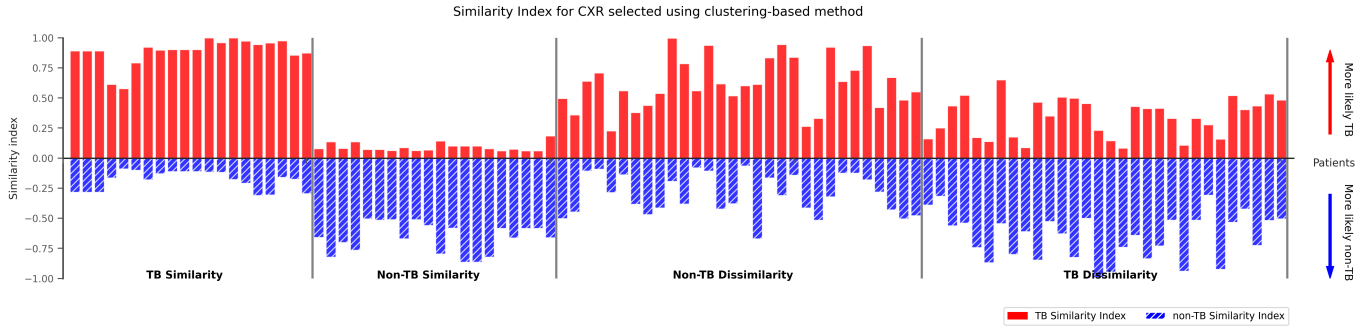


Figure 11: Neighborhood similarity index for the selected CXR images. The x-axis represents the images sorted selection group and y-axis the normalized measure.

Images sampled by the similarity criteria are still located in regions with homogeneous characteristics. In contrast, the images sampled by the dissimilarity criteria are distributed in hybrid areas with similar probabilities for both classes (TB and not TB) of the dataset (with a small the advantage for the opposite class of the image).

The last step of the analysis is to assess whether the images would be sampled using the validation set of the newly created models using the same criteria based on selecting samples with the higher silhouette indices. Compared with the previous analysis, Only 54,5% would be selected. If we consider only the top 25 samples, only 6 images would be selected. Focusing only on the dissimilar criteria, no previously selected images would be selected. This is a result of the fact that the images are not being used in the training process. Although clusters and neighborhoods were preserved mostly stables through different models, their centroids were slightly shifted compared to the ones on the original map. Also, the centroids moved in an opposite direction than the projected image vectors, so, even if they are active neurons in the central region of the cluster, they are not with the highest values for the silhouette index.

6 Conclusion

This paper described how images from a popular CXR dataset were sampled using a machine-learning pipeline. In the end, 100 images were chosen after extensive exploitation of the non-linear relations among the entire dataset due to the deployment feature engineering, data projection, and clustering techniques working in conjunction.

The images were sampled according to similarity/dissimilarity criteria and the silhouette index was used as a measure of the quality of the selection. The highest the silhouette index, the higher the chance of picking an image with the characteristics observed through the visual inspection of the obtained clusters. The obtained images maximized the representativeness of the dataset and this was demonstrated using exemplary images to illustrate.

SOM models were used to extract the topological structure of the dataset and the neighborhood characteristics were considered for selecting the images. Hence, a neighborhood similarity index was proposed for the measurement of how similar a given image is related to its neighbors. The index computes the empirical probability for each class being represented within the SOM unit by considering the number of occurrences of each class in the own neuron and its neighbors. The index was able to measure the similarity and dissimilarity of the images.

After the sampling task, a cross-validation procedure was performed in order to evaluate whether the operational model would be able to preserve the topological structure for new datasets. Results show that the newly created models had the structural aspects from the original map and the clusters were similar on both methods. Also, sampled images didn't migrate from one cluster to another and they were preserved in similar neighborhoods. However, the sampled images wouldn't be picked up by the generalization model, as the centroids of the clusters shift away from the projected validation data. Maybe CXR images could be sampled using the neighborhood similarity index, as it seems a more robust measure for this selection criteria. Further studies are needed to investigate this possibility.

Regardless, future research could continue to explore the use of the proposed sampling technique for other public CXR datasets, like India dataset [50] and Montgomery County dataset [47].

References

[1] J. Chakaya, M. Khan, F. Ntoumi, E. Aklillu, R. Fatima, P. Mwaba, N. Kapata, S. Mfinanga, S. E. Hasnain, P. D. Katoto *et al.*. "Global Tuberculosis Report 2020–Reflections on the Global TB burden, treatment and prevention efforts". *International Journal of Infectious Diseases*, vol. 113, pp. S7–S12, 2021.

[2] W. H. Organization. "WHO consolidated guidelines on tuberculosis", 2021.

[3] A. J. Zimmer, J. S. Klinton, C. Oga-Omenka, P. Heitkamp, C. N. Nyirenda, J. Furin and M. Pai. "Tuberculosis in times of COVID-19". *J Epidemiol Community Health*, vol. 76, no. 3, pp. 310–316, 2022.

[4] A. C. Nachiappan, K. Rahbar, X. Shi, E. S. Guy, E. J. Mortani Barbosa Jr, G. S. Shroff, D. Ocazionez, A. E. Schlesinger, S. I. Katz and M. M. Hammer. "Pulmonary tuberculosis: role of radiology in diagnosis and management". *Radiographics*, vol. 37, no. 1, pp. 52–72, 2017.

[5] J. Burrill, C. J. Williams, G. Bain, G. Conder, A. L. Hine and R. R. Misra. "Tuberculosis: A Radiologic Review". *RadioGraphics*, vol. 27, no. 5, pp. 1255–1273, 2007. PMID: 17848689.

[6] N. Kumar, S. Bhargava, C. Agrawal, K. George, P. Karki and D. Baral. "Chest radiographs and their reliability in the diagnosis of tuberculosis." *Journal of the Nepal Medical Association*, vol. 44, no. 160, 2005.

[7] W. H. Organization *et al.*. "Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches". Technical report, World Health Organization, 2016.

[8] C. Dye, P. Glaziou, K. Floyd and M. Raviglione. "Prospects for tuberculosis elimination". *Annual review of public health*, vol. 34, 2013.

[9] S. Kulkarni and S. Jha. "Artificial intelligence, radiology, and tuberculosis: a review". *Academic radiology*, vol. 27, no. 1, pp. 71–75, 2020.

[10] M. Harris, A. Qi, L. Jeagal, N. Torabi, D. Menzies, A. Korobitsyn, M. Pai, R. R. Nathavitharana and F. Ahmad Khan. "A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis". *PloS one*, vol. 14, no. 9, pp. e0221339, 2019.

[11] E. Kotter and M. Langer. "Computer aided detection and diagnosis in radiology". *European Radiology*, vol. 21, 2011.

[12] T. Pande, C. Cohen, M. Pai and F. Ahmad Khan. "Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review". *The International Journal of Tuberculosis and Lung Disease*, vol. 20, no. 9, pp. 1226–1230, 2016.

[13] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, M. A. Ayari and M. E. H. Chowdhury. "Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization". *IEEE Access*, vol. 8, pp. 191586–191601, 2020.

[14] Q. H. Nguyen, B. P. Nguyen, S. D. Dao, B. Unnikrishnan, R. Dhingra, S. R. Ravichandran, S. Satpathy, P. N. Raja and M. C. H. Chua. "Deep Learning Models for Tuberculosis Detection from Chest X-ray Images". In *2019 26th International Conference on Telecommunications (ICT)*, pp. 381–385, 2019.

[15] S. Hwang, H.-E. Kim, J. Jeong and H.-J. Kim. "A novel approach for tuberculosis screening based on deep convolutional neural networks". In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, pp. 750–757. SPIE, 2016.

[16] S. Jaeger, A. Karargyris, S. Candemir, J. Siegelman, L. Folio, S. Antani and G. Thoma. "Automatic screening for tuberculosis in chest radiographs: a survey". *Quantitative imaging in medicine and surgery*, vol. 3, no. 2, pp. 89, 2013.

[17] J. Burt, N. Torosdagli and N. Khosravan. "Deep learning beyond cats and dogs: recent advances in diagnosing brest cancer with deep neural networks". *The British Journal of Radiology*, vol. 91, 2018.

[18] H.-P. Chan, R. K. Samala and L. M. Hadjiiski. "CAD and AI for breast cancer - recent development and challenges". *The British Journal of Radiology*, vol. 93, 2019.

[19] J. M. Lew, C. Mao, M. Shukla, A. Warren, R. Will, D. Kuznetsov, I. Xenarios, B. D. Robertson, S. V. Gordon, D. Schnappinger *et al.*. "Database resources for the tuberculosis community". *Tuberculosis*, vol. 93, no. 1, pp. 12–17, 2013.

[20] C. Qin, D. Yao, Y. Shi and Z. Song. "Computer-aided detection in chest radiography based on artificial intelligence: a survey". *Biomedical engineering online*, vol. 17, no. 1, pp. 1–23, 2018.

[21] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer *et al.*. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[22] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] T. Kohonen. "The self-organizing map". *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[24] M. Srinivas, D. Roy and C. K. Mohan. "Discriminative feature extraction from X-ray images using deep convolutional neural networks". In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 917–921. IEEE, 2016.

[25] J. Guérin and B. Boots. "Improving image clustering with multiple pretrained cnn feature extractors". *arXiv preprint arXiv:1807.07760*, 2018.

[26] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken and C. I. Sánchez. "A survey on deep learning in medical image analysis". *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[27] M. Ayaz, F. Shaukat and G. Raja. "Ensemble learning based automatic detection of tuberculosis in chest x-ray images using hybrid feature descriptors". *Physical and Engineering Sciences in Medicine*, vol. 44, no. 1, pp. 183–194, 2021.

[28] M. Oloko-Oba and S. Viriri. "Tuberculosis abnormality detection in chest X-rays: a deep learning approach". In *International Conference on Computer Vision and Graphics*, pp. 121–132. Springer, 2020.

[29] S. S. Meraj, R. Yaakob, A. Azman, S. Rum, A. Shahrel, A. Nazri and N. F. Zakaria. "Detection of pulmonary tuberculosis manifestation in chest X-rays using different convolutional neural network (CNN) models". *Int. J. Eng. Adv. Technol.(IJEAT)*, vol. 9, no. 1, pp. 2270–2275, 2019.

[30] B. Oltu, S. Güney, B. Dengiz and M. Ağıldere. "Automated Tuberculosis Detection Using Pre-Trained CNN and SVM". In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 92–95. IEEE, 2021.

[31] P. Anu Priya and E. Vimina. "Tuberculosis Detection from CXR: An Approach Using Transfer Learning with Various CNN Architectures". In *International Conference on Communication, Computing and Electronics Systems*, pp. 407–418. Springer, 2021.

[32] F. Corona, M. Mulas, R. Baratti and J. A. Romagnoli. "On the topological modeling and analysis of industrial process data using the SOM". *Computers & Chemical Engineering*, vol. 34, no. 12, pp. 2022–2032, 2010.

[33] I. Jolliffe and J. Cadima. "Principal component analysis: a review and recent developments". *Philos Trans A Math Phys Eng Sci.*, vol. 374, 2016.

[34] S. Kaski. *Self-Organizing Maps*, pp. 886–888. Springer US, Boston, MA, 2010.

[35] K.-L. Du. "Clustering: A neural network approach". *Neural networks*, vol. 23, no. 1, pp. 89–107, 2010.

[36] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[37] P. Rousseeuw. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[38] R. Mayer, R. Neumayer, D. Baum and A. Rauber. "Analytic comparison of self-organising maps". In *International Workshop on Self-Organizing Maps*, pp. 182–190. Springer, 2009.

[39] R. Ramon-Gonen and R. Gelbard. "Cluster evolution analysis: Identification and detection of similar clusters and migration patterns". *Expert Systems with Applications*, vol. 83, pp. 363–378, 2017.

[40] S. Bag, S. K. Kumar and M. K. Tiwari. "An efficient recommendation generation using relevant Jaccard similarity". *Information Sciences*, vol. 483, pp. 53–64, 2019.

[41] C. Liu, Y. Cao, M. Alcantara, B. Liu, M. Brunette, J. Peinado and W. Curioso. "TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network". *Proc. - Int. Conf. Image Process. ICIP*, vol. 2017-Septe, pp. 2314–2318, 2018.

[42] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization", 2017.

[43] M. Ojala and G. C. Garriga. "Permutation Tests for Studying Classifier Performance". *Journal of Machine Learning Research*, pp. 1833–1863, 2010.

[44] T. Ciodaro, D. Deva, J. de Seixas and D. Damazio. "Online particle detection with Neural Networks based on topological calorimetry". In *Journal of Physics: Conference Series*, volume 368, p. 012030. IOP Publishing, 2012.

[45] D. Schmidt. "A Survey of Singular Value Decomposition Methods for Distributed Tall/Skinny Data". In *2020 IEEE/ACM 11th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA)*, pp. 27–34. IEEE, 2020.

[46] G. Vettigli. "MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map", 2018.

[47] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu and G. Thoma. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases". *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, pp. 475, 2014.

[48] K. He, X. Zhang, S. Ren and J. Sun. "Deep Residual Learning for Image Recognition". *CoRR*, vol. abs/1512.03385, 2015.

[49] T. Kohonen. "Exploration of very large databases by self-organizing maps". In *Proceedings of international conference on neural networks (icnn'97)*, volume 1, pp. PL1–PL6. IEEE, 1997.

[50] A. Chauhan, D. Chauhan and C. Rout. "Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation". *PloS one*, vol. 9, no. 11, pp. e112980, 2014.