Machine Learning - Final Project Report

Using Segmentation and Classification Algorithms To detect Tuberculosis in Pulmonary Chest X-Rays Given data from Montgomery(USA) and Shenzhen(China)

Alexandre Landeau
CentraleSupélec
b00396898@essec.edu

Caroline Favart
CentraleSupélec
b00705417@essec.edu

Juan Londono
CentraleSupélec
b00605470@essec.edu

Vivien Robert
CentraleSupélec
b00531125@essec.edu

ABSTRACT

The purpose of our project is to successfully predict tuberculosis based solely on Chest X-Rays (CXR). Tuberculosis is the deadliest single infectious agent disease, and it mostly affects countries lacking radiological expertise. Models able to predict tuberculosis by evaluating CXR images might be a powerful way to tackle this issue.

To do so, we have at our disposal a dataset of 800 images from Montgomery, USA (138 images) and from Shenzhen, China (662 images). Those images have different dimensions, and a preparatory phase was necessary to be able to apply the modelling techniques.

We have set our focus on a machine learning approach, even though models based on state-of-the-art deep learning techniques are more performant. Indeed, our approach allows us to better understand our findings since it is based on feature extraction and description for each image. Those features can for instance be plotted over the original image. We then perform a visual bag-of-words technique on the different predictors, and finally use classic classification algorithms to predict whether a patient is sick or not.

We used Accuracy and AUC indicators to evaluate our models and to compare them to different studies aimed at solving the same problem. We achieved top scores when compared to other machine learning techniques, but our results are still largely outperformed by deep learning programs.

ACM Reference format:

Alexandre Landeau, Caroline Favart, Juan Londono and Vivien Robert. 2018. Using Segmentation and Classification Algorithms To detect Tuberculosis in Pulmonary Chest X-Rays Given data from Montgomery(USA) and Shenzhen(China). In *Proceedings of ACM, 2 pages.* https://doi.org/10.1145/1234567890

KEYWORDS

Feature: pixel significant in an image, corresponding a point where the image gradient changes suddenly, on an edge or a corner for example.

Feature detector: algorithm used to detect features, the input is an image and the output a list of feature coordinates. We used four industry standard feature detectors: SIFT, ORB, SURF, FAST.

Feature descriptor: algorithm used to describe features. The list of criteria the algorithm uses is the same for every feature, allowing to compare features based on those common criteria.

Bag of words: Natural Language Processing (NLP) technique used for document classification based on the recurrence of words from same clusters (similar to lexical fields) in each document

INTRODUCTION/MOTIVATION

Tuberculosis (TB) is to this day the deadliest single infectious agent, above HIV/AIDS. It is estimated that it caused more than 1.3 million deaths in 2017, and that 10 million new persons were infected during the same year (1). More to the point, the 6.4 million cases reported represented only 64% of the estimated 10.0 million new cases that occurred in 2017 (1). Such a gap can be explained by a mixture of underreporting of detected cases and underdiagnoses.

Chest X-ray is a rapid imaging technique that allows lung abnormalities to be identified. CXR is used to diagnose conditions of the thoracic cavity, including the airways, ribs, lungs, heart and diaphragm.

Although being one of the most efficient methods for detecting TB, the CXR technique was previously used at the latter stages of TB

screenings, mostly due to high costs and availability problems. Indeed, TB mostly affects developing countries, with limited access to radiologists (2).

However, CXR has lately been promoted by the World Health Organization (WHO) as a useful tool in the fight against the disease. The WHO now recommends the use of CXR at the early stages of detection, helping detect asymptomatic people in the early course of the disease.

As operating costs for CXR have decreased during the past decade, new Machine and Deep Learning tools are emerging, and could prove to be a powerful way to compensate for the lack of radiological expertise in developing countries. Machine Learning solutions are already available on the market, but the research on this field is ongoing, as the performance of the models can still be improved (3,4).

Model performance will play a key role in the development of these new solutions and on the generalization of CXR as a major way to diagnose TB. We aim at tackling this issue by implementing Segmentation and Classification methods, in the hope of finding a strong and reliable model.

PROBLEM DEFINITION

1 Objective

Given 2 datasets of anterior-posterior pulmonary x-ray images (CXR), we would like to create a model to detect Tuberculosis (TB) amongst the patients with a high accuracy rate. This means that we want to maximize the number of True Positives while minimizing the number of False Negatives.

This problem is a classification problem based on computer vision. Our methodology will be composed of the following main steps (which will be further describe in the Methodology part):

- Features engineering: we will perform different types of features detection algorithms to extract the coordinate of points of interest (especially, the tuberculosis scars) in the images. Then, we will use different descriptor extraction algorithm to build vectors describing the detected points of interest.
- Bag of visual words: we will use NLP process and clustering algorithms on the created vectors to generate a "common vocabulary" and project each image on this vocabulary.
- Classification Model training: we will then apply different supervised machine learning models of different complexities to compare them and find the best one

2 Datasets

We have two different anterior-posterior pulmonary x-ray images (CXR) datasets, one Chinese and one American:

- the Chinese "Shenzhen set" (662 observations): consists in a standard digital image database for Tuberculosis, created by the National Library of Medicine, Maryland, USA in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China and is composed of Shenzhen out-patient clinics data.
- the American "Montgomery set" (138 observations): acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA.

The study especially focuses on machine learning that gains knowledge through supervised learning. Supervised learning, implies that an expert is present to define correct and incorrect responses. In our case, the datasets were previously analysed by medical experts so that for each CXR we have metadata and clinical readings. This information has been used by previous researchers to label each observation as Tuberculous case or not. For the "Shenzhen set", we have 336 cases with manifestation of tuberculosis and 326 normal cases. The "Montgomery set" contains 138 posterior-anterior x-rays, of which 80 x-rays are normal and 58 x-rays are abnormal with manifestations of tuberculosis. In total, we have a combined dataset of 800 observations with 394 Tuberculosis cases and 406 normal cases.

At last, previous researchers also provided manually-made masks for each dataset to better target the interesting part of each CXR image and delete noises. Yet, the masks and the formatting of the CXRs were heterogeneous across the 2 datasets at the beginning. We decided to perform pre-processing techniques to modify them and build a homogenous dataset.

RELATED WORK

Various methods have been used to tackle our issue. Amongst those we can cite:

1. Machine Learning (ML) approaches using a combination of textural abnormality and shape detection.
2. ML approaches using lung segmentation, texture and shape feature extraction, and classification with support vector machines.
3. Deep Learning (DL) techniques.

The first method corresponds to the CAD4TB, the only commercially available software. It obtains similar results than clinical officers and radiologists (10), with an AUC ranging from

0.71 to 0.84 in five studies (3). As we wanted our model to outperform traditional detection methods, we decided to set this model as our Null model.

We then had to choose between the second ML method, and DL.

Although most of the recent work related to our topic uses state-of-the-art deep learning techniques for image classification, we decided to focus on the second method.

Indeed, as opposed to DL, this ML method works more as a "white bow". The texture and shape feature extraction allow us to know what the different features detected by our algorithm are, for each image. After assigning descriptors to each feature, we can plot the original image alongside its different features.

This gives us great insights on how our algorithm works, which would have been hard to get with DL methods. We sacrificed some accuracy to gain in interpretability.

The methods we use are based on the work of D.Lowe, University of British Columbia, published in 2004: Distinctive Image Features from Scale-Invariant Keypoints. He introduced the Scale Invariant Feature Transform (SIFT) algorithm, which extract keypoints and compute its descriptors (11). His new method allowed to detect corners and edges regardless of scale.

This was the starting point of the ML techniques we are interested on. Many other algorithms were developed based on that. Some to improve the feature extraction part, such as SURF or FAST (12) algorithms, and some to improve feature description. We will specially set focus on the BRIEF algorithm introduced in 2010 (13).

Our aim was to replicate the work previously done in the field, by testing and combining the different algorithms published throughout the years. Another publication based on the same datasets presented an accuracy of 84% maximum (4). This was therefore our objective, keeping in mind that we could easily have obtained more than 90% accuracy using DL methods.

METHODOLOGY

1 Data preparation

The first part of our work was data preparation. Given the fact that we were focused on lungs, we wanted to have filters on every image so that the model only focuses on lungs. We also needed images of same size to train the model. However, all images were not the same size, nor resolution or even orientation. We also only had filters for the Montgomery dataset, meaning 138 images out of our 800 X-rays. Given these differences we treated the two datasets separately:

For the Montgomery set:

- as the filters of left and right lungs were in different files, we first merged them and then dilated them by 15 pixels to get smoother edges
- we resized every x-ray and filter into a 512*512 pixels image so that the model input has a constant size
- we plotted the results for improved control on the process

For the Shenzhen set:

- we started by using the 662 manual filters created by Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Peng Gang, Wei Zeng, Yuri Gordienko for their study on the Shenzhen dataset
- we smoothed every filter by dilating them by 15 pixels
- as x-rays and filters were in portrait or landscape orientation we resized them all into a 512*512 pixels image so that the model input has a constant size
- we plotted the results for improved control on the process

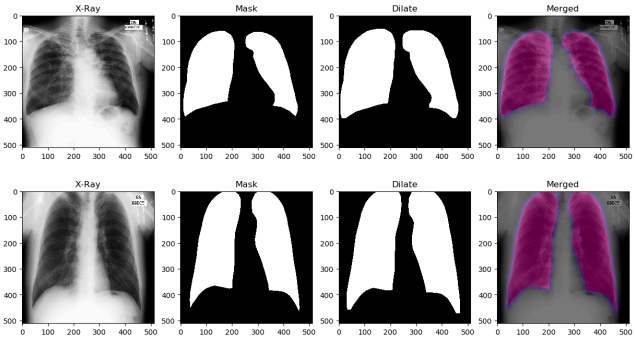

Figure 1: Train and Test x-rays from the Montgomery Set, with comparison between original and dilated masks

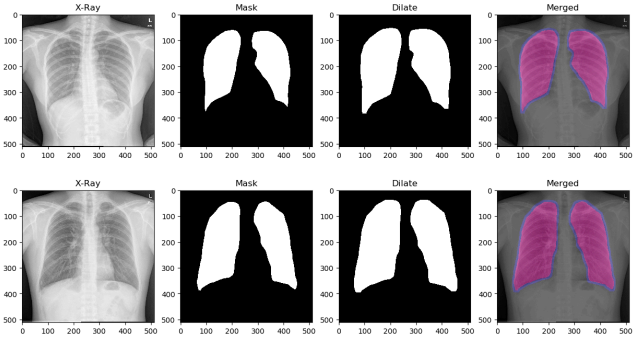

Figure 2: Train and Test x-rays from the Shenzhen Set, with comparison between original and dilated masks

2 Feature Engineering

The next part consists in Feature Engineering. For every image we must find significant features to describe the image instead of running the model on raw images. We used descriptor extraction, meaning we found areas of interest such as edges or corners and described them using various methods. Each method is based on two steps:

- feature detection: which winds which pixels are interesting in the image and returns their coordinates
- descriptor extraction: which describes every feature using the same descriptors, the descriptors are proper to each model and each feature is transformed into a vector of its coordinates regarding each descriptor

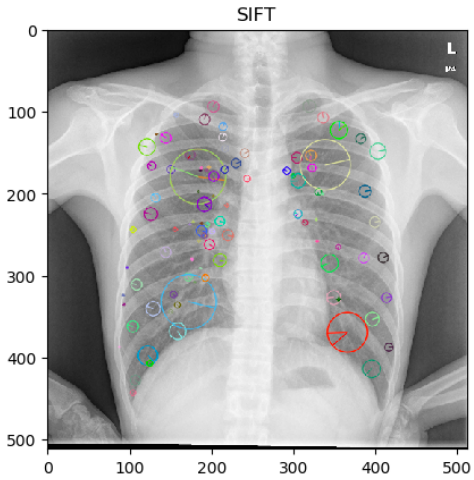

Figure 3: Representation of the predictors and descriptors for an x-ray using SIFT

As an example, if we use the SIFT method, SIFT returns for an image a matrix of size m x 128 where m is the number of features detected for the image and 128 is the number of descriptors SIFT uses. So, if our dataset is composed of n images, our data after feature extraction will look like the matrix below:

$$\begin{bmatrix} features_0 \\ features_1 \\ .... \\ .... \\ .... \\ features_n \end{bmatrix}$$

$$where\ features_i\ \ is\ \ a\ \ array\ of\ \ \ dimension\ \ m \times 128$$

Figure 4: Schematic representation of the dataset after feature extraction

3 Bag of visual words

Our images are now described by a various number of descriptors, all of them having the same size. The third step in our methodology is to apply a NLP methodology to the descriptors, the bag of words, to classify the images. We consider every image as a document, each descriptor being a “visual word”.

In order to sort the “documents”, we need to sort visual words into visual lexical fields, we used k means clustering to generate a common vocabulary, consisting in all the centroids of the k means. Then we used vector quantization to project each image on the vocabulary, transforming all images into histograms of k bins. Now every image is transformed into a numerical vector of length k, each coordinate being the projection of the image on a word of the common vocabulary. The problem can now be treated as a regular machine learning problem, we just scale the histograms before switching to the model training part.

4 Model Training

The problem is now a classification problem with k numerical dimensions. The last part is the model training and tuning. Having scaled all the predictors, we decided to compare different types of classification models (support vectors, clusters, ensemble methods), we chose the following models:

- Naive Bayes Classifier
- k Nearest Neighbours Classifier
- Support Vector Classifier
- AdaBoost Classifier
- Random Forest Classifier
- Extreme Gradient Boosting Classifier

To optimize the models, we tuned them using Grid Search and cross validation, following a method detailed in the Evaluation part of the report.

We finally estimated the generalization error by using a test set detailed in the Evaluation part as well.

5 Limitations

As the model uses centroids to cluster the predictors we don't have choice when it comes to using different algorithms for the segmentation part. Indeed, k means is the standard clustering method which uses centroids and other methods, such as Spectral Clustering, don't return centroids and are then unusable in this method.

When it comes to overfitting, we see that all models tend to underfit a little bit, this could explain why we have better performances with extractors that return higher number of descriptors, meaning which have higher variance, such as Brief and Sift. This could also explain why convNets tend to outperform our models, as they have the possibility to embed much more features.

EVALUATION

1 Creation of a test set

After the data preparation, we isolated 50 American X-Rays and 50 Shenzhen X-rays in a test set, with their respective masks and dilated images. We put the rest of the dataset in the training set and used it for training our models. We chose to take half of each data origin in the test set, to reduce the bias in the evaluation that our model might have, as we trained it on more Chinese cases.

2 Model tuning

We performed model tuning at each step of the above-mentioned pipeline, to obtain the best suited model.

Indeed, we first tried out different thresholds for the feature extractors to compare the model performances, for instance on the STAR extractor (which we finally dismissed because the FAST extractor performed better), we manually adjusted the response Threshold.

Then, for each method of feature extraction, we tried different k for k-means. It allowed us to plot the Sum of Squared Errors corresponding to each k (in range 1 to 100) and to use the Elbow-Curve method to define k to optimize the clustering.

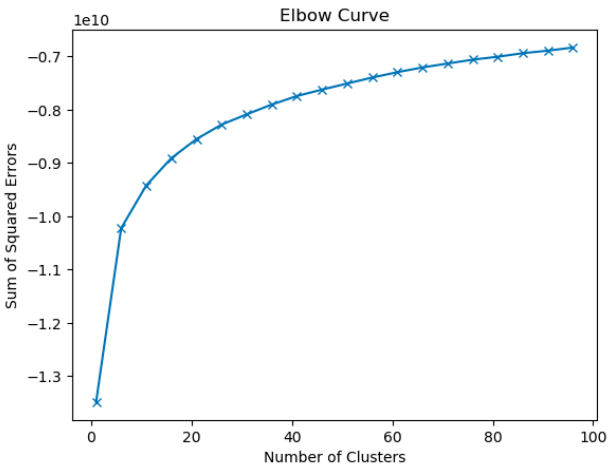

Figure 5: Elbow Curve for the SIFT features

Nonetheless, we observed that the value of k optimizing the clustering and given by the elbow method was not the value of k maximizing the accuracy. This was mentioned by the founders of the Bag of Words method, who decided to take the value of k maximizing the final performances. Following the same logic, we took the value of k maximizing the test accuracy.

Finally, we also optimized the model parameters of each classifier, using the function GridSearchCV.

3 Model evaluation

To evaluate our models, we used three different methods:

1) We computed the cross-validation score as an intermediary step.

Cross-validation	SVC	kNN	Multinomial Naïve Bayes	Adaboost	XGB
SIFT	0.74	0.74	0.76	0.75	0.75
ORB	0.75	0.74	0.76	0.75	0.73
SURF	0.79		0.74		0.77
Fast + BRIEF		0.79	0.75	0.78	0.80

Figure 6: computed Cross-validation score for each model

2) We computed the accuracy of each model on the test set, which we used to select the best model.

Figure 7 represents, for each feature extraction and description method, the accuracy of each classification algorithm.

Accuracy	SVC	kNN	Random Forest	Multinomial Naïve Bayes	Adaboost	XGB
SIFT	0.77	0.77	0.74	0.58	0.74	0.58
ORB	0.73	0.7	0.76	0.7	0.72	0.68
SURF	0.79		0.75	0.79		0.76
Fast + BRIEF	0.85	0.79	0.81	0.75	0.82	0.83

Figure 7: computed Accuracy for each model

It is interesting to see that for each feature extractor, the best-performing classifiers are not the same: the Random Forest classifier performs best for ORB, whereas SVC gives the best results for SIFT, ORB and BRIEF.

We can also mention that there is no overfitting, since the test accuracy values are in the same range as the cross-validation scores performed on the training set. Only the Multinomial Naïve Bayes seems to overfit for SIFT, as well as the XGB for SIFT and ORB. We could almost say that the other models are underfitting, as we have seen that the accuracy increases when the number of features increases.

Moreover, we observe that the combination of the FAST extractor and BRIEF descriptor enhances the performance of the classifiers, with a best score of 85% of accuracy for an SVC. Thus, we would choose this model.

3) We computed the AUC values of each model.

AUC	SVC	kNN	Random Forest	Multinomial Naïve Bayes	Adaboost	XGB
SIFT	0.77	0.73	0.69	0.61	0.73	0.72
ORB	0.70	0.71	0.71	0.65	0.70	0.74
SURF	0.78		0.72	0.68		0.76
Fast + BRIEF	0.84	0.77	0.81	0.73	0.81	0.81

Figure 8: computed AUC for each model

We observe that the values change a bit, for ORB, XGBoost becomes the best classifier for instance. But the best solution remains to use FAST for feature extraction, BRIEF for feature description and SVC for the classification.

CONCLUSIONS

Our final results match the original objectives, as our best model yields an 85% accuracy rate, and an 0.84 AUC:

- We have outperformed our Null model, the CAD4TB commercial solution and the clinical officers' and radiologists' predictions, which both yield AUCs around 0.71 and 0.84. This alone is an achievement, as if we were to release our model today, it would have a real impact in actual tuberculosis predictions.

- We have obtained the same accuracy than the authors of the article Automatic tuberculosis screening using chest radiographs (4) published in 2014. As this team uses the same datasets and similar methods, this result was to be expected, and was one of our aims.

To improve the quality of our predictions, testing the model on datasets from different countries might be of real interest. Indeed, our data was mainly composed of Chinese patients and predictions might vary according to local specificities.

It was highly interesting to test different image classification techniques through machine learning methods that allowed us to visualize and better understand our findings.

However, we are well aware that such techniques might never find real-word applications. Indeed, best performing models are now all based on deep learning techniques and can achieve up to 98% accuracy. As models' predictive properties are the main priority in the medical field, it would be best suited to choose accuracy over interpretability in the face of such trade-off.

REFERENCES

Data sources :

[1] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[2] Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China

Publications:

[1] World Health Organization. Global tuberculosis report 2018. Sep 18, 2018.
[2] World Health Organization. Chest radiography in tuberculosis detection. 2016.
[3] Paras Lakhani, Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Thomas Jefferson University Hospital, Apr 2017. https://pubs.rsna.org/doi/10.1148/radiol.2017162326
[4] Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Xue Z, Palaniappan K, Singh RK, Antani S, Thoma G, Wang YX, Lu PX, McDonald CJ. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging , Feb 2014.
[5] Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, Karargyris A, Antani S, Thoma G, McDonald CJ. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Med Imaging, Feb 2014.
[6] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, Cédric Bray, Visual Categorization with Bags of Keypoints, ResearchGate, 2014 https://www.scribd.com/document/300446906/Visual-categorization-with-bags-of-keypoints
[7] Kushal Vyas, Bag of Visual Words model for Image Classification and Recognition https://kushalvyas.github.io/BOV.html
[8] Ian London, Image Classification in Python with Visual Bag of Words https://ianlondon.github.io/blog/visual-bag-of-words/

[9] Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Peng Gang, Wei Zeng, Yuri Gordienko, Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation, ELANO, 2018 https://arxiv.org/abs/1803.01199

[10] Maduskar, P.; Muyoyeta, M.; Ayles, H.; Hogeweg, L.; Peters-Bax, L.; van Ginneken, B. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. The International Journal of Tuberculosis and Lung Disease, Volume 17, Number 12, 1 December 2013

[11] D.Lowe. Distinctive Image Features from Scale-Invariant Keypoints. University of British Columbia, 2004.

[12] Edward Rosten and Tom Drummond, "Machine learning for high speed corner detection" in 9th European Conference on Computer Vision, vol. 1, 2006, pp. 430–443.

[13] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "BRIEF: Binary Robust Independent Elementary Features", 11th European Conference on Computer Vision (ECCV), Heraklion, Crete. LNCS Springer, September 2010.