J Med Syst (2018) 42: 146
https://doi.org/10.1007/s10916-018-0991-9

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs

Szilárd Vajda[1] · Alexandros Karargyris[2] · Stefan Jaeger[4] · K.C. Santosh[3] · Sema Candemir[4] · Zhiyun Xue[4] · Sameer Antani[4] · George Thoma[4]

Received: 9 October 2017 / Accepted: 12 June 2018 / Published online: 29 June 2018

Abstract

To detect pulmonary abnormalities such as Tuberculosis (TB), an automatic analysis and classification of chest radiographs can be used as a reliable alternative to more sophisticated and technologically demanding methods (e.g. culture or sputum smear analysis). In target areas like Kenya TB is highly prevalent and often co-occurring with HIV combined with low resources and limited medical assistance. In these regions an automatic screening system can provide a cost-effective solution for a large rural population. Our completely automatic TB screening system is processing the incoming CXRs (chest X-ray) by applying image preprocessing techniques to enhance the image quality followed by an adaptive segmentation based on model selection. The delineated lung regions are described by a multitude of image features. These characteristics are than optimized by a feature selection strategy to provide the best description for the classifier, which will later decide if the analyzed image is normal or abnormal. Our goal is to find the optimal feature set from a larger pool of generic image features, –used originally for problems such as object detection, image retrieval, etc. For performance evaluation measures such as under the curve (AUC) and accuracy (ACC) were considered. Using a neural network classifier on two publicly available data collections, –namely the Montgomery and the Shenzhen dataset, we achieved the maximum area under the curve and accuracy of 0.99 and 97.03%, respectively. Further, we compared our results with existing state-of-the-art systems and to radiologists' decision.

Keywords Tuberculosis · Chest x-ray · Automatic chest x-ray analysis · Feature selection · Neural networks · HOG · Automatic TB screening

Introduction

Tuberculosis (TB) – according to the 2017 WHO report [41], is considered one of the major life threats beside HIV (human immunodeficiency virus), with a mortality rate of 1.3 million people among the 10.4 million people developing the disease each year. Cure rates over 90% have been described in clinical studies. However, it still remains a major challenge due to the presence of TB in tandem with HIV in 1.7 million cases out of the reported 10.4 million

https://ceb.nlm.nih.gov/repos/chestImages.php

This article is part of the Topical Collection on *Advanced Computational Intelligence and Soft Computing in Medical Imaging*

✉ Szilárd Vajda
szilard.vajda@cwu.edu

Alexandros Karargyris
akararg@us.ibm.com

Stefan Jaeger
stefan.jaeger@nih.gov

K.C. Santosh
Santosh.KC@usd.edu

Sema Candemir
sema.candemir@nih.gov

Zhiyun Xue
zhiyun.xue@nih.gov

Sameer Antani
sameer.antani@nih.gov

George Thoma
george.thoma@nih.gov

[1] Central Washington University, Ellensburg, WA, USA

[2] IBM Almaden Research, San Jose, CA, USA

[3] University of South Dakota, Vermillion, SD, USA

[4] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ones. Among the population contracting the virus, 90% are adults, 65% are male, and 56% are coming from only five countries: Indonesia, Pakistan, India, the Philippines and China.

TB is an infectious disease caused by the bacillus Mycobacterium tuberculosis, which typically affects the lungs. It spreads through the air when people with active TB cough, sneeze, or otherwise expel infectious bacteria [52]. TB is most prevalent in Africa and Southeast-Asia, where widespread poverty and malnutrition reduces resistance to the disease. The most common method for diagnosing TB worldwide is sputum smear microscopy (developed more than 100 years ago), in which bacteria are observed in sputum samples examined under a microscope. Following recent developments in TB diagnostics, the use of rapid molecular tests for the diagnosis of TB and drug-resistant TB is increasing, as highlighted in WHO's reports [40, 41]. In countries with more developed laboratory capacity TB cases are also diagnosed via culture methods (the current reference standard). However, these methods are currently rather expensive, and not easily applicable in low-resourced regions such as Africa. In these areas chest X-ray (CXR) is still the most prominent TB detection method in use.

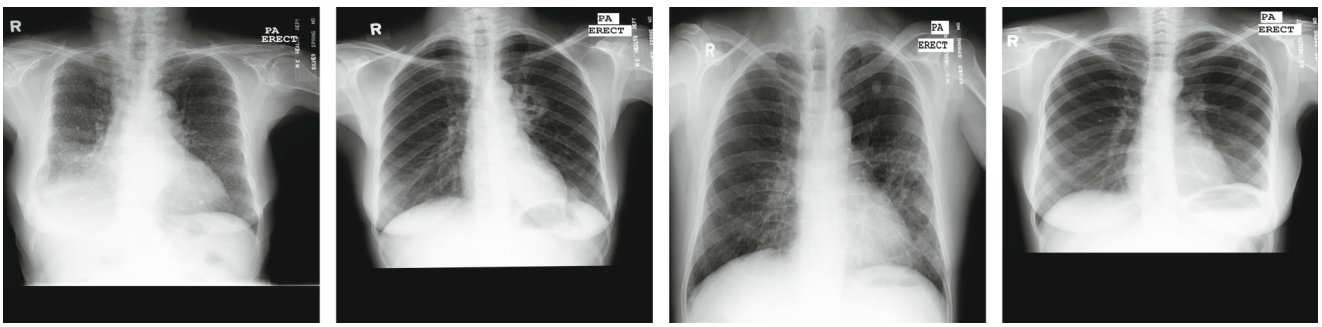
Tuberculosis is exhibited in CXR images in form of cavitations, consolidations, infiltrates, blunted costophrenic angles, opacities, pleural effusion and thickening, pneumonia, horizontal fissure displacement, hilar enlargement and small broadly distributed nodules [52], among other radiological manifestations. These changes can often be detected in CXRs in the form of corrupted and/or deformed lung profiles [27], disruptions in the lung shapes, intensity changes in the lung tissue [23], texture abnormalities [8], etc. Some prominent TB manifestations can be observed in Fig. 1. Besides the design and development of a deployable and reliable CXR screening system, our major aim is to select the best and complementing features. These specially selected characteristics will help the underlying classifier to produce a complex decision surface necessary to distinguish normal CXRs from the abnormal ones.

Our objective with the feature selection [17] implemented for our CXRs classification scenario was three-fold: *i)* improve the prediction performance of the underlying classifier, *ii)* provide an optimal feature set suitable to describe abnormalities such as TB in the lung, and *iii)* provide a direct comparison of our results with those published by Jaeger et al. [23]. In addition to the main goal to find an optimal feature set providing high classification accuracy, our secondary goal was to select a fast and well performing classifier such as an artificial neural network [3, 53]. Such a network is able to define complex non-linear decision surfaces necessary to distinguish TB cases from non-TB cases relying only on features. The feature selection will also make possible an overall shorter processing time due to the fact that only a reduced number of features is to be extracted and used in the classification process.

In this paper, we propose an end-to-end system capable of detecting different lung abnormalities from CXRs analysis using only image processing and machine learning. The rest of the paper is structured as follows: "Related work" gives a brief overview of the state-of-the-art, "Methods" discusses the methods in use, – involving lung segmentation, features description, features selection and classification. "Experiments" provides a brief description of the used chest X-ray collections, the evaluation protocols and the different results. Finally, a brief summary highlighting the strengths of our paper is provided in "Conclusion".

Related work

Recently, we note an increased focus on automatic chest radiography [2, 11, 24, 25, 28, 29, 31, 44] due to the more affordable prices for X-ray machines, and the huge potential residing in the automatic image processing [16]. Such tools analyze these digital images without any external human involvement [30]. Even though, in the last few years many papers have been published in computer-aided diagnosis



(a) Miliary nodules (b) Cavitary infiltrates (c) Infiltrates (d) Pleural effusion

Fig. 1 Different Tuberculosis manifestations in chest X-Ray images

(CAD) targeting chest x-ray images [26], there are only a limited number of systems which can accurately read chest radiographs [14, 31].

Due to the uncontested success of deep convolutional neural networks, in the recent years different works appeared in the medical image analysis field [2, 22, 28, 29, 31]. Instead of using the traditional feature extractions followed by classification, the researchers in this new paradigm tasked the networks to extract automatically [18] the separating characteristics from X-rays, MRIs, etc. However, such methodologies need very large training samples [21] and some small deformations like calcifications or infiltrates might not be detected properly [2]. Do to this fact we focused our current research on the classical solution, where well defined image characteristics can describe the different abnormalities and is usable for reduced size data too.

Nodule detection is becoming one of the popular research focuses, due to the very well defined aspect of the problem. Even some commercial systems are available on the market [46] helping radiologist to localize and diagnose lung cancer [19]. However, nodules are one of many representations of TB, besides consolidations, infiltrates, blunted costophrenic angles, opacities, pleural effusion and thickening, pneumonia, horizontal fissure displacement, or hilar enlargement. Due to the high complexity of the problem to detect these different type of TB manifestations, recent studies concentrate more on specific topics, such as lung segmentation [5, 8], temporal subtraction for bone scans [47], or other aspects such as detection of catheters and pneumothorax, texture analysis or shape analysis [15].

To overcome human involvement in lung cancer detection addressed by the seminal work of Lodwick et al. [33] by converting the visual images of roentgenograms into numerical sequences, the current research shifted the focus to more sophisticated and automatic feature extraction methods. These features are able to describe the different phenomenons encountered in the different CXRs. Van Ginneken et al. [15] identified these possible features as being texture related as patterns are diffuse. The analysis of pixel neighborhood intensities can reveal certain specific characteristics. As the authors mention it, is hard for a radiologist ”to get a clue” why these image features relate to certain diseases. However, to mimic the radiologists' reading habits, computer scientists should transcribe the reading knowledge in a more formal way. The extraction of all types of image characteristics (intensity, shape profiles, wavelets, etc.) should be followed by feeding these characteristics into sophisticated classifiers such as neural networks, Support Vector Machines (SVM), etc. The noise – caused by the image acquisition or size of the lung region of the analyzed subject, etc. [36] can skew the results. To reduce these type of artifacts, different methods such as rib segmentation [1], rib supression or histogram equalization [23] have been implemented. The perceptual errors committed by human readers can be corrected with focused analysis using systematic search strategies, coning devices, etc.

Depeursing et al. [10] proposed a study to compare different classification methods involving five different classifiers applied to three types of feature groups: gray-level histograms, air components and quincunx wavelet frame coefficients with B-spline wavelets. Similar attempts have been proposed by Jaeger et al. [23] involving classifiers such as SVM, multi-layer perceptron, decision tree and linear regression. In both cases SVMs provided the best performances. The work [54] presents a rather small scale experiment (77 images) only for nodule detection involving feature extraction. The features – mainly intensity values, wavelets, Gabor coefficients, multi-scale Hurst features, etc, in total 67 different characteristics were selected using a genetic algorithm (GA) by minimizing the overall classification error. With the method they managed to reduce the features number to 25. However, there is no direct comparison showing the importance of the feature selection.

In general, there is no clear understanding why some features perform better than others and there is no clear view how those image features can actively contribute in the classification. Therefore, a clear understanding of the features and their combination is a necessity in order to provide a well defined framework in the future for pulmonary disease detection and classification.

Methods

This section describes the different processing steps of the system: starting with lung segmentation using atlas-based segmentation, the feature selection, and finally the classification which provides the user with a confidence measure for each analyzed image belonging to the normal or abnormal cases.

Lung segmentation

In our system, we use an atlas-based lung segmentation algorithm. The atlas is a set of CXRs from several patients and their expert delineated lung boundaries. The system first chooses the most similar models to the patient X-ray by measuring the lung shape similarities. Then, it warps the selected models to the patient X-ray using a registration algorithm. This algorithm uses the scale invariant feature transform (SIFT) flow (i.e., SIFT-flow) registration approach [32], which computes the corresponding pixels of image pairs according to their SIFT feature similarity. The average of the registered models will constitute the patient-specific lung model. The system then

combines the CXR intensity values and lung model with an objective function to decide for the final boundary. The segmentation solves the objective function with a graph-cut energy minimization approach [4].

The system produces state-of-the-art results on a public set (c.f., JSRT set [45]) reaching 0.954 ± 0.0015 coverage. Similar scores were reported for the Montgomery collection, where 0.941 ± 0.034 coverage was obtained. For more details about this stage, we refer to the work by Candemir et al. [5].

Features description

To characterize normal and TB suspicious CXR lung segments, we considered three different feature sets. The feature *Set A* is inspired from object detection [16, 37] and was used with success in a previous work [23]. The feature *Set B* has been utilized with success by Rahman et al. [42] for a medical CBIR system. Finally, we considered basic shape features, which can also be powerful to characterize abnormalities. For pleural effusion the lower part of the lung is not visible due to the accumulated fluid in the thoracic cavity, thus producing a blunt costophrenic angle [35] and a considerably modified lung shape [52].

All features have only been extracted from the lung regions detected by the atlas-based segmentation method (see "Lung segmentation") preceded by a histogram equalization to enhance the overall contrast of the analyzed CXR images.

Set A: Is a versatile and compact feature set combining shape, edge and texture descriptors. The final feature representation is built by concatenating the different descriptors (histograms) extracted from the segmented lung regions. In particular, the following shape and texture descriptors were considered: Intensity Histogram (IH), Gradient Magnitude Histogram (GM), Shape Descriptor Histogram (SD), Curvature Descriptor Histogram (CD), Histogram of Oriented Gradient (HOG) [9], Local Binary Pattern (LBP) [39]. A modified multiscale approach proposed by Frangi et al. [13] is considered to compute the eigenvalues of Hessian matrix needed for the shape and curvature descriptors. The Hessian describes the second-order surface curvature properties of the local image intensity surface. The normalization makes these descriptors intensity invariant. Jaeger et al. [23] determined that quantizing these features into 32 bins provides good discrimination performance. The size of the feature descriptor is 192.

Set B: Is a rather diversified and low-level feature collection involving intensity, edge, texture, color and shape moment features. The final feature representation is built by concatenating the different descriptors (histograms) extracted from the segmented lung regions. In particular, the following descriptors were considered: Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD) from MPEG-7 standard [34], Color and Edge Direction Descriptor (CEDD) [6], Fuzzy Color and Texture (FCTH) [7], Tamura texture descriptor, Gabor texture feature [20], and other texture features such as primitive length (PL), edge frequency (EF), and autocorrelation (AC) [48]. This feature set is larger, comprising 595 features.

Set C: Is a focused feature collection involving only shape measurements calculated from the lung shapes provided by the standard MATLAB® implementation. For our purpose size, orientation, eccentricity, extent, centroid, and bounding box were considered. The dimension of this feature set is 12. For each lung segment 6 different features were extracted and later concatenated. For details please refer to the help provided by MATLAB's *regionprops*[1].

Set C contains only similar types of features, while set A and B are a mixture of all sorts of features, as they were used separately for different pattern recognition tasks [16, 37, 42]. Therefore, we have not seen the necessity to classify the features based on their properties and nature.

Feature selection and classification

In many systems devoted to better CXR analysis [8, 14, 23, 24, 27], the authors do not specifically motivate their selection for the particular features in use. Rather, they just borrow well-known features from image processing [16]. Without any specific motivation, – excepting color, edges or texture, which are applicable to all kinds of object detection tasks [37], content based image retrieval (CBIR) [42] works do not consider particularly crafted features to characterize abnormalities such as TB. While some features can complement each other, – by improving the discriminating power of the descriptor [50], some features might work in the detriment of others, thus the selection of features from a larger pool is necessary and useful for further consideration.

For our purpose we considered a wrapper type feature selection model [43, 49]. Instead of aiming to reach a certain accuracy level, – often used as selection criterion, we conducted an exhaustive search among the different feature combinations. Given n different features, the number of possible combinations is:

$$N = \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n - 1 \quad (1)$$

[1] http://www.mathworks.com/help/images/ref/regionprops.html

Due to the different feature combinations, the corresponding feature subsets were concatenated, and a train/test procedure launched to established the best combination. The feature combination providing the maximum ACC or AUC was retained as the winner combination. Due to the exponential nature of the experiment, a parallel approach of the wrapper method was implemented. The selection/training/testing phases for each particular feature combination is completely independent so such parallelism was possible. We are aware that this exhaustive search might take longer than a random search looking for a certain threshold, but due to the fact, that this search has to be performed only once, we preferred performance maximization over speed.

For classification, a neural network-based classifier was considered. Our choice for the selection of the classifier might sound a bit arbitrary, but we conducted some preliminary experiments using Support Vector Machine (SVM). The results were not encouraging, and they were in the range of the original experiments conducted by Jaeger et al. [23]. Neural networks in particular are known for their capacities of estimating complex decision surfaces [3] and handling multi-class problems. Due to the large numbers of features to be handled (up to 799 dimensions), and the lack of information about the possible correlations among the different feature components, a fully connected multi-layer perceptron network was utilized. The number of neurons in the input layer was selected based on the dimensionality of the input feature vector. The number of output neurons was also set based on the possible outcomes: normal and abnormal. The number of neurons in the hidden layer was estimated based on several trial runs. Finally, for the experiments 10 neurons were considered as being optimal in the hidden layer. For training, error-backpropagation strategy was considered, while for *learning rate* $= 0.001$, and *momentum* $= 0.2$ were used. The different parameters were established based on several trial runs. For the number of hidden neurons in the hidden layer, we considered the criteria to have as less possible neurons to keep the complexity low as possible, therefore the recognition time becomes faster. More sophisticated network optimizations involve pruning algorithms [51, 53] could be considered, however this kind of optimization goes beyond the scope of the current research.

Experiments

This section gives a detailed description of the data in use, the evaluation protocols considered for the experiments. Finally, the different experiment setups are described followed by some comparisons.

Data

For the experiments two different, publicly available CXR data collections were considered. The images in these studies were de-identified by the data providers, and are exempted from IRB review at their institution. The data was exempted from IRB review (No. 5357) by the NIH office of Human Research Protection Program. The Montgomery dataset, – a representative subset of a larger image repository, was collected over many years within the tuberculosis control program of the Department of Health and Human Services of Montgomery County. The set contains 138 posteroanterior CXRs, among which 80 CXRs are normal ones, while the remaining 58 CXRs are abnormal cases (presenting some sort of abnormality indicating TB). The Shenzhen dataset is from Shenzhen No. 3 Hospital (Shenzhen, China), one of the largest hospitals in China for infectious diseases. The CXR images belong to outpatient clinics. The collection contains 342 normal, and 334 abnormal cases. For more details about the data, please refer to [23].

Evaluation protocols

In order to properly evaluate the performance of the current system, several measures were considered. Accuracy (ACC) and area under the curve (AUC) were selected [12] to measure these performances. We considered these measurements and not others such as MCC (Matthew correlation coefficients), because we wanted to directly compare our results with the results presented in [23]. Beside the ACC, the AUC is a necessary measure to understand the behavior of the underlying classifier. Due to the special application in question, namely deciding if a specific CXR contains abnormalities, there is a high need to control the true positive rate, as nobody should be missed if his/her CXR contains a certain type of abnormality [38]. The ROC curve also gives us the possibility to adjust our classification threshold for the purpose of our application. Each of our experiments follows a 10x cross-validation protocol. The reported results are the average scores of the different folds.

Results

Different experiments were conducted involving the Montgomery, and the Shenzhen collection, respectively. First some results are reported using as input for the classifier the feature set A, feature set B, and feature set C, respectively. The goal of these experiments was to show the importance of these features separately, as well as their strength by combining them together.

In Table 1. is to be observed the discriminative power of the feature set A, involving less features than set B. The intensity histogram, the local binary patterns, the histogram of oriented gradients, etc. seem to be more powerful than the features borrowed from MPEG-7 standard. Similar trend is to be observed in the work proposed by Jaeger et al. [23]. The increased scores in our cases suggest also the fact, that the neural network is able to better estimate the decision surface than a support vector machine (SVM).

While the first column in the table show individual results for the different feature sets, the last column involves all the features described in detail in "Features description". This extended feature set is focusing more on the common representational effort of these features –by stitching them together. Apparently, the combined feature set involving set A, B, and C can not overcome the individual results generated by set A, because set B and C are introducing a certain number on confusions.

Due to the the limited number of blunt costophrenic angle appearances in the analyzed collections, the shape features have only limited description power. The majority of the CXRs available in our collections have TB manifestations inside the lung regions, and not that much along the boundaries usually involving severe shape deformations. However, shape features (Set C) can be still considered as a reliable source to separate abnormal CXRs from normal ones, when the TB manifestation is to be observed on the shape such as pleural effusion [35, 52].

In Table 2. similar conditions were considered as in the case of Table 1., but this time instead of measuring ACC, the AUC is measured to show the real strength of the neural classifier, –by varying the threshold applied to the accuracy. The results achieved for the feature set A are very promising, achieving almost perfect scores for the Shenzhen data (AUC = 0.99), and promising score (AUC = 0.87) for the Montgomery collection. These scores provide a real proof that it is possible to set up a classifier which provides almost perfect classification rates. Following the trend discussed earlier, the feature set B and feature set C provide moderate results, due to their limitation describing and capturing the specific shape and orientations of lung, ribs, etc. The ROC curves for the Montgomery and Shenzhen collections, involving set A are shown in Fig. 2.

However, these characteristics are classical image features used in object recognition or content based image retrieval. Nobody analyzed their particular contribution to the final recognition. Therefore, our feature selection experiment identified some 17 different features belonging to Set A(#6), B(#10), and C(#1). The experiments in Tables 3 and 4. show those optimized feature sets which provide the highest accuracy, and maximized area under the curve, respectively. Our optimization criteria for the best selection of features was $max\{ACC\}$ and $max\{AUC\}$.

Table 1 Accuracy (ACC) measures reported for the different feature representations for different data collections

Dataset	Set A(%)	Set B(%)	Set C(%)	Set {A,B,C}(%)
Montgomery	78.30	72.47	65.82	69.45
Shenzhen	95.57	81.06	70.40	92.00

Table 2 Area under the curve (AUC) measures reported for the different feature representations for different data collections

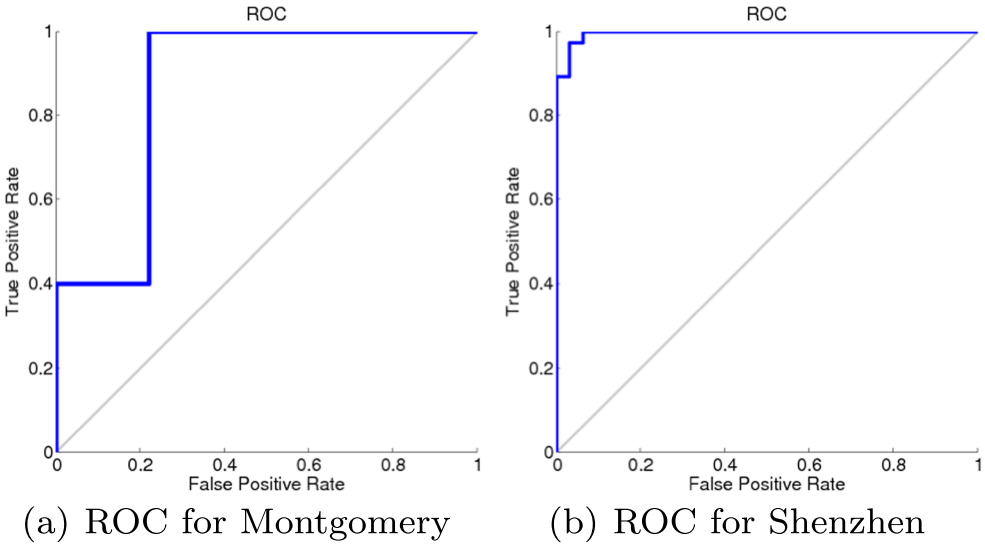
Dataset	Set A	Set B	Set C	Set {A,B,C}
Montgomery	0.87	0.72	0.71	0.79
Shenzhen	0.99	0.90	0.77	0.97

For feature selection each possible feature combination was trained/tested on a 10x cross-validation basis, and the average scores were reported. The scores for the optimized feature collections (see Tables 3 and 4.) are way more accurate than the results obtained by the original features (see Tables 1 and 2.) obtaining a net ACC gain of 6.45% (Montgomery) and 1.46% (Shenzen), respectively. The AUC net gain goes up to 4% (Montgomery), while for the Shenzhen data the same AUC has been achieved, more precisely 0.99, – a result which is acceptable considering the importance of the correct classification of the true positive cases (abnormalities).

Is to be noted that for the feature selection, – due to the nature of the evaluation protocol (no dedicated training/test set), all the folds contributed to the selection of the best feature collection, therefore the results could be biased [49].

In order to support the correctness of our choice for the selection, we published the standard deviation (σ) for the results coming from the 10 folds. While the standard deviation is low for the results reported for the Shenzhen collection (see Table 4.), the spread is really high for the Montgomery collection (see Table 3.). The rather high σ level for the second collection can be explained with the relatively low number of CXRs and the unbalanced aspect of this collection. The standard deviation in this case can be considered as a measure to see how far we would be in case of a dedicated test set to test with.

While the results reported for the feature selection might be biased, not only is the accuracy gain substantial, but all features need much more time to be extracted, thus influencing the overall processing time of each chest radiograph. Beside the increased accuracy/area under the curve, we also would like to focus our discussion towards the selected features. Among the selected ones is to be observed a net dominance of features such as IH, CD, LBP, HOG belonging to the original feature set



(a) ROC for Montgomery (b) ROC for Shenzhen

Fig. 2 ROC curve for the Montgomery and Shenzhen collection using the feature set A for classification

A. From the set B, features like FTCH, GLCM, Gabor and EF were considered, showing the strength of these particular features in the overall evaluation. As one can see, these features describe texture, intensity, edginess, etc., properties valuable to distinguish normal and abnormal CXRs. Far beyond the comparison to the baseline system [23], our ultimate goal is to find the best possible feature combination, and deploy the system in Kenya to accurately detect TB positive patients. Therefore, for direct evaluation our feature selection scores are somehow biased, but for the upcoming chest x-ray images to be analyzed, we discovered the best feature descriptors to be considered.

The optimization also supports the fact that shape features do not contribute to the classification, and their usage should be rather considered in a pre-filtering phase, before starting a thorough analysis of the radiograph. Such a filter beside a costophrenic angle estimator [35] could help to quickly identify lung shape abnormalities which are strong indicators of different types of lung diseases.

As our screening application is to be used in a TB prevalent area such as Kenya, it is important to provide the healthcare providers with a reliable tool for screening, – avoiding any misclassification of the abnormal cases. While it is still acceptable to identify a healthy lung as being abnormal, none of the abnormal cases should be missed. Therefore, we show some results in Table 5. for both collections involving recall values of 0.90, 0.95, and 0.99, respectively.

The false positive rate (FPR) for the Shenzhen collection is rather promising and acceptable. The results produced for the Montgomery collection are more modest. One possible explanation could be the reduced number of samples present in the collection. It is known that for statistical classifiers such as neural networks, to adjust the different weights through the learning process a multitude of different samples is necessary [3]. This condition is more fulfilled for the Shenzhen collection, where the number of samples is over 300, both for abnormal and normal cases, respectively.

To directly compare our results, we considered the system proposed by Jaeger et al. [23]. In this paper the authors are focusing exactly on the same data and using same type of experiments as we provided in this current paper. For quality measurements accuracy (ACC) and area under the curve (AUC) were considered, both being adequate measures to decide about the quality of the system. While the authors of the previously mentioned work report different type of results, for the sake of clarity, only their

Table 3 Results for the optimized feature set involving the Montgomery collections

Optimization	Result	σ	Selected features
$max\{ACC\}$	84.75%	11.16	{FTCH,EF,IH,GM,SD,CD,LBP}
$max\{AUC\}$	0.91	0. 11	{FTCH,GLCM,EF,IH,SD,CD}

Table 4 Results for the optimized feature set involving the Shenzhen collections

Optimization	Result	σ	Selected features
$max\{ACC\}$	97.03%	1.71	{CLD,Gabor,GLCM,EF,IH,HOG,LBP}
$max\{AUC\}$	0.99	0.005	{Gabor,EF,GM,HOG,LBP}

Table 5 False positive rates for recall values of 0.90, 0.95 and 0.99 for Montgomery and the Shenzhen collections

Dataset	0.90	0.95	0.99
Montgomery	0.261	0.261	0.261
Shenzhen	0.003	0.011	0.062

best scores will be mentioned in the comparison. To our best knowledge, there are no other works reporting results on the Montgomery and Shenzhen benchmark collections. As one can see in Table 6, the improvements both ACC (11.47%) and AUC (11%) reported by our system are rather significant for the Shenzhen collection. The improvement of the AUC with 1% for the Montgomery collection is statistically insignificant. The comparison with the achieved scores by the feature selection (see Tables 3 and 4.) would be even more impressive than the results obtained by the feature set A (see Tables 1 and 2.). However, such comparison would not be exactly accurate.

We also compared the performance of our system with human reading performances. For that purpose we used the results reported by Jaeger et al. [23]. For the experiments two independent radiologists were asked to read the CXRs belonging to the Montgomery collection. This process was completely independent from our experiments, and it was based only on visual inspection of the frontal chest X-rays. In Table 7. a detailed confusion matrix is presented, showing how human readers perform in classifying the CXRs into normal and abnormal cases, respectively. By calculating the accuracy (ACC) obtained by the radiologists, one can observe the fact, that the accuracy (81.86%) achieved by the radiologist is still higher than the accuracy (78.30%) reported by our system, considering the exact same conditions for the Montgomery collection. With more specific features and more training samples available, we are confident that the scores provided by the automatic systems will increase gradually. All these results point us to the conclusion that automatic screening systems are necessary and helpful. With the corresponding medical expertise provided by the radiologist, machines can also classify with high accuracy and reliability chest radiographs for the benefit of the overall diagnostic process.

Table 6 ACC and AUC comparisons between the results reported by Jaeger et al. [23] and the results produced by our system (see Our)

Dataset	ACC [23]	ACC (Our)	AUC [23]	AUC (Our)
Montgomery	78.3%	78.3%	0.86	0.87
Shenzhen	84.10%	95.57%	0.88	0.99

Table 7 Comparison of human consensus performance with ground truth of Montgomery collection [23]

		Consensus		
		+	–	
Ground truth	+	58	0	58
	–	25	55	80
		83	55	138

Conclusion

In this paper, we presented a completely automatic frontal chest radiograph screening system able to detect healthy lungs and spot abnormal ones - carrying different type of Tuberculosis manifestations. Due to the focus group specificities (Kenya's rural population), – involving limited resources and limited medical personnel, the development of such mobile screening systems is important, and it has a huge benefit for the public health endeavors sustained currently in Kenya.

Our main goal, besides the description of the automatic CXR screening system, was to gain a deeper understanding; why some features can carry the necessary information to separate the abnormal cases from the normal cases using and some others do not possess such capability. The majority of the current systems just borrow some well-known features from the literature, –considered for larger purpose object detection or content-based image retrieval, and apply a classification scheme on top of that. Our solution provides a wrapper based feature selection to find a particular feature combination which minimizes the classification error rate, and maximizes the area under the curve.

Considering three different feature sets involved in a previous study, we managed to select, –in a data-driven manner, those particular feature combinations which maximize the overall performances of the classification systems for the different CXR collections. Among the selected features we can enumerate features such as Gabor, Fuzzy Color and Texture Histogram, Intensity Histogram, Shape Descriptor, Local Binary Pattern, Curvature descriptor, Histogram of Oriented Gradient, Edge Frequency, features which can be considered in the future for similar classification tasks. These characteristics are more concentrated on the overall image quality, edginess and texture, –properties which can apparently distinguish between normal and abnormal CXRs. However, we are aware that these results are reported for only two different frontal chest X-ray collections, namely the Montgomery and Shenzhen collections. These publicly available collections contain only a limited number of X-rays, but beside our main goal to detect important and descriptive features from a larger collection, we

also wanted to provide a direct comparison of our results with those published by Jaeger et al. [23], – hence the choice of the data.

Our classification shows a net improvement of up to 11.45% accuracy and 11% improvement in the area under the curve for the Shenzhen collection. Considering the results involving the feature selection, the scores can go even higher. Admittedly, our feature selection scheme is biased, however, with this selection we managed to identify the feature subset on which the trained upcoming deployed system in Kenya could provide the best recognition score.

Our experiments involving false positives rates for fixed recall values of 0.90, 0.95, and 0.99 show that we can define such a threshold mechanism based on the ROC analysis which could provide high specificity values. This is a necessity for such medical applications.

To further improve the automatic part in the classification process, one could extract automatically features from the analyzed lung regions using an encoder type network. Combining both type of features could lead to increased performances. Beside concentrating on the features some special attention can be focused on the classifier too. Instead of using one classifier to identify all sort of TB manifestations, specialized classifiers could better identify certain particular anomalies such as infiltrates, calcifications, pleural effusion, etc.

Besides identifying the normal cases the precise detection of abnormal cases could be deferred to other, more sophisticated healthcare facilities such as hospitals or clinics where more in-depth investigations can take place. The comparison of our results with medical experts' readings shows that automatic systems such as ours can be considered in the screening process. Such computer-aided diagnosis systems can work side-by-side with medical experts providing a second opinion and actively helping pulmonary diagnosis of patients.

Acknowledgments This research is supported in past by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine, and Lister Hill National Center for Biomedical Communications (LHNCBC).

The authors are grateful to Mr. Rodney Long for the fruitful discussions during the development of this project.

Funding Information This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

Compliance with Ethical Standards

Conflict of interests Authors declare that they have no conflict of interest.

Ethical approval All images used in this study were collected prior to this study during routine clinical care. They were de-identified at source and have been exempted from review (NIH IRB# 5357).

References

1. Banik, S., Rangayyan, R. M., and Boag, G. S., Automatic segmentation of the ribs, the vertebral column, and the spinal canal in pediatric computed tomographic images. *J. Digit. Imaging* 23(3):301–322, 2010.
2. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., and Greenspan, H., Chest pathology detection using deep learning with non-medical training. In: 12Th IEEE International Symposium on Biomedical Imaging, ISBI 2015, brooklyn, April 16-19, 2015, pp. 294–297. https://doi.org/10.1109/ISBI.2015.7163871, 2015.
3. Bishop, C. M., *Neural networks for pattern recognition*. New York: Oxford University Press, inc., 1995.
4. Boykov, Y., Veksler, O., and Zabih, R., Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(11):1222–1239, 2001.
5. Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G. R., and McDonald, C. J., Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* 33(2):577–590, 2014.
6. Chatzichristofis, S. A., and Boutalis, Y. S., Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: *Proceedings of the 6th International Conference on Computer Vision Systems, ICVS'08*, pp. 312–322. Berlin: Springer, 2008.
7. Chatzichristofis, S. A., and Boutalis, Y. S., Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In: *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '08*, pp. 191–196. Washington: IEEE Computer Society, 2008.
8. Chauhan, A., Chauhan, D., and Rout, C., Role of Gist and PHOG Features in Computer-Aided Diagnosis of Tuberculosis without Segmentation. PLoS ONE 9(11): e112980. https://doi.org/10.1371/journal.pone.0112980, 2014.
9. Dalal, N., and Triggs, B., Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20–26 june 2005, San diego, pp. 886–893, 2005.
10. Depeursinge, A., Iavindrasana, J., Hidki, A., Cohen, G., Geissbühler, A., Platon, A., Poletti, P., and Müller, H., Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *J. Digit. Imaging* 23(1):18–30, 2010.
11. Doi, K., Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput. Med. Imaging Graph.* 31(4–5):198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002, 2007. http://www.sciencedirect.com/science/article/pii/S0895611107000262. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.
12. Fawcett, T., An introduction to ROC analysis. *Pattern Recogn. Lett.* 27(8):861–874, 2006.
13. Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A., Muliscale vessel enhancement filtering. In: Medical Image Computing and Computer-assisted Intervention - MICCAI'98, first international conference, Cambridge, October 11-13, 1998, pp. 130–137, 1998.
14. van Ginneken, B., ter Haar Romeny, B. M., and Viergever, M. A., Computer-aided diagnosis in chest radiography: A survey. *IEEE Trans. Med. Imaging* 20(12):1228–1241, 2001.
15. van Ginneken, B., Hogeweg, L., and Prokop, M., Computer-aided diagnosis in chest radiography: Beyond nodules. *Eur. J. Radiol.* 72(2):226–230. https://doi.org/10.1016/j.ejrad.2009.05.061, 2009.

http://www.sciencedirect.com/science/article/pii/S0720048X09003581. Digital Radiography.
16. Gonzalez, R. C., and Woods, R. E., *Digital image processing*. 3 ed. Upper Saddle River: Prentice-Hall, Inc., 2006.
17. Guyon, I., and Elisseeff, A., An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3:1157–1182, 2003. http://dl.acm.org/citation.cfm?id=944919.944968.
18. Hinton, G., and Salakhutdinov, R., Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507, 2006.
19. de Hoop, B., Schaefer-Prokop, C., Gietema, H. A., de Jong, P. A., van Ginneken, B., van Klaveren, R. J., and Prokop, M., Screening for lung cancer with digital chest radiography: Sensitivity and number of secondary work-up ct examinations. *Radiology* 255(2):629–637, 2010.
20. Howarth, P., Yavlinsky, A., Heesch, D., and Ruger, S., Medical image retrieval using texture, locality and colour. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., and Magnini, B. (Eds.) *Multilingual Information Access for Text, Speech and Images, Lecture Notes in Computer Science*, Vol. 3491, pp. 740–749. Berlin: Springer, 2005.
21. Hwang, S., Kim, H., Jeong, J., and Kim, H., A novel approach for tuberculosis screening based on deep convolutional neural networks. In: Medical imaging 2016: Computer-aided diagnosis, San diego, 27 february - 3 march 2016, p. 97852w, 2016.
22. Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K., Abnormality detection and localization in chest x-rays using deep convolutional neural networks. CoRR arXiv:1705.09850, 2017.
23. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F. M., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S., Thoma, G. R., Wang, Y., Lu, P., and McDonald, C. J., Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* 33(2):233–245, 2014.
24. Jaeger, S., Karargyris, A., Candemir, S., Siegelman, J., Folio, L., Antani, S., and Thoma, G., Automatic screening for tuberculosis in chest radiographs: a survey. *Quant. Imaging Med. Surg.* 3(2):89, 2013.
25. Karargyris, A., Siegelman, J., Tzortzis, D., Jaeger, S., Candemir, S., Xue, Z., Santosh, K. C., Vajda, S., Antani, S. K., Folio, L., and Thoma, G. R., Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays. *Int. J. Comput. Assist. Radiol. Surg.* 11(1):99–106, 2016. https://doi.org/10.1007/s11548-015-1242-x.
26. Katsuragawa, S., and Doi, K., Computer-aided diagnosis in chest radiography. *Comput. Med. Imaging Graph.* 31(4–5):212–223. https://doi.org/10.1016/j.compmedimag.2007.02.003, 2007. http://www.sciencedirect.com/science/article/pii/S0895611107000286. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.
27. KC, S., Vajda, S., Antani, S., and Thoma, G., Automatic pulmonary abnormality screening using thoracic edge map. In: Int. Symposium on computer-based medical systems, pp. 360–361, 2015.
28. Kim, H. E., and Hwang, S., Scale-invariant feature learning using deconvolutional neural networks for weakly-supervised semantic segmentation. CoRR arXiv:1602.04984, 2016.
29. Kooi, T., Litjens, G. J. S., van Ginneken, B., Gubern-mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., and Karssemeijer, N., Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* 35:303–312. https://doi.org/10.1016/j.media.2016.07.007, 2017.
30. Li, Q., Recent progress in computer-aided diagnosis of lung nodules on thin-section {CT}. *Comput. Med. Imaging Graph.* 31(4–5):248–257. https://doi.org/10.1016/j.compmedimag.2007.02.005, 2007. http://www.sciencedirect.com/science/article/pii/S0895611107000316. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.
31. Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I., A survey on deep learning in medical image analysis. *Med. Image Anal.* 42:60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005.
32. Liu, C., Yuen, J., and Torralba, A., Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(5):978–994, 2011.
33. Lodwick, G. S., Keats, T. E., and Dorst, J. P., The coding of roentgen images for computer analysis as applied to lung cancer. *Radiology* 81(2):185–200, 1963.
34. Lux, M., Caliph & emir: Mpeg-7 photo annotation and retrieval. In: *Proceedings of the 17th ACM International Conference on Multimedia, MM '09*, pp. 925–926. New York: ACM, 2009.
35. Maduskar, P., Hogeweg, L., Philipsen, R., and van Ginneken, B., 2013.
36. McAdams, H. P., Samei, E., James Dobbins, I., Tourassi, G. D., and Ravin, C. E., Recent advances in chest radiography. *Radiology* 241(3):663–683, 2006.
37. Murphy, K. P., Torralba, A., Eaton, D., and Freeman, W. T., Object detection and localization using local and global features. In: Toward Category-level Object Recognition, pp. 382–400, 2006.
38. Obuchowski, N. A., Roc analysis. *Fundamentals of Clinical Research for Radiologists* 184(2):364–372, 2005.
39. Ojala, T., Pietikäinen, M., and Harwood, D., A comparative study of texture measures with classification based on featured distributions. *Pattern Recogn.* 29(1):51–59, 1996.
40. Organization, W. H., Global tuberculosis report. http://apps.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf. Online; accessed 23-March-2015, 2012.
41. Organization, W. H., Global tuberculosis report. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf. Online; accessed 20-April-2018, 2017.
42. Rahman, M. M., You, D., Simpson, M. S., Antani, S., Demner-fushman, D., and Thoma, G. R., Interactive cross and multimodal biomedical image retrieval based on automatic region-of-interest (ROI) identification and classification. *IJMIR* 3(3):131–146, 2014.
43. Saeys, Y., Inza, I. N., and Larrañaga, P., A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517, 2007.
44. Santosh, K. C., Vajda, S., Antani, S. K., and Thoma, G. R., Edge map analysis in chest x-rays for automatic pulmonary abnormality screening. *Int. J. Comput. Assist. Radiol. Surg.* 11(9):1637–1646. https://doi.org/10.1007/s11548-016-1359-6, 2016.
45. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., Matsui, M., Fujita, H., Kodera, Y., and Doi, K., Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists detection of pulmonary nodules. *Am. J. Roentgenol.* 174:71–74, 2000.
46. Shiraishi, J., Li, F., and Doi, K., Computer-aided diagnosis for improved detection of lung nodules by use of posterior-anterior and lateral chest radiographs. *Acad. Radiol.* 14(1):28–37. https://doi.org/10.1016/j.acra.2006.09.057, 2007. http://www.sciencedirect.com/science/article/pii/S1076633206005599.
47. Shiraishi, J., Li, Q., Appelbaum, D., and Doi, K., Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin. Nucl. Med.* 41(6):449–462. https://doi.org/10.1053/j.semnuclmed.2011.06.004, 2011. http://www.sciencedirect.com/science/article/pii/S0001299811000742. Image Perception in Nuclear Medicine.
48. Singh, S., and Sharma, M., Texture analysis experiments with meastex and vistex benchmarks. In: Singh, S., Murshed, N., and Kropatsch, W. (Eds.) *Advances in Pattern Recognition — ICAPR*

2001, Lecture Notes in Computer Science, pp. 419–426. Berlin: Springer, 2001.
49. Smialowski, P., Frishman, D., and Kramer, S., Pitfalls of supervised feature selection. *Bioinformatics* 26(3):440–443, 2010.
50. Vajda, S., Rangoni, Y., and Cecotti, H., Semi-automatic ground truth generation using unsupervised clustering and limited manual labeling: Application to handwritten character recognition. *Pattern Recogn. Lett.* 58(0):23–28, 2015.
51. Wang, S. H., Muhammad, K., Lv, Y., Sui, Y., Han, L., and Zhang, Y. D., Identification of alcoholism based on wavelet renyi entropy and three-segment encoded jaya algorithm. *Complexity* 2018:13, 2018.
52. Weinberger, S., Cockrill, B., and Mandel, J., Principles of pulmonary medicine. Elsevier Health Sciences, 2013.
53. Zhang, Y., Sun, Y., Phillips, P., Liu, G., Zhou, X., and Wang, S., A multilayer perceptron based smart pathological brain detection system by fractional fourier entropy. *J. Med. Syst.* 40(7):1–11, 2016.
54. Zhu, Y., Tan, Y., Hua, Y., Wang, M., Zhang, G., and Zhang, J., Feature selection and performance evaluation of support vector machine (svm)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *J. Digit. Imaging* 23(1):51–65, 2010.