CHAPTER

Chest Radiograph Pathology Categorization via Transfer Learning

13

Idit Diamant*,[1], Yaniv Bar*,[1], Ofer Geva*, Lior Wolf*, Gali Zimmerman*, Sivan Lieberman†, Eli Konen†, Hayit Greenspan*

*Tel-Aviv University, Ramat-Aviv, Israel** *Sheba Medical Center, Tel-Hashomer, Israel†*

CHAPTER OUTLINE

CHAPTER POINTS

- Overview of X-ray analysis: from BoW to deep learning
- Deep learning can be used via transfer learning from an existing network
- Medical images can be represented via deep-network signature
- Transfer learning enables image multi-label categorization

[1]Equal contributors.

Deep Learning for Medical Image Analysis. DOI: 10.1016/B978-0-12-810408-8.00018-3

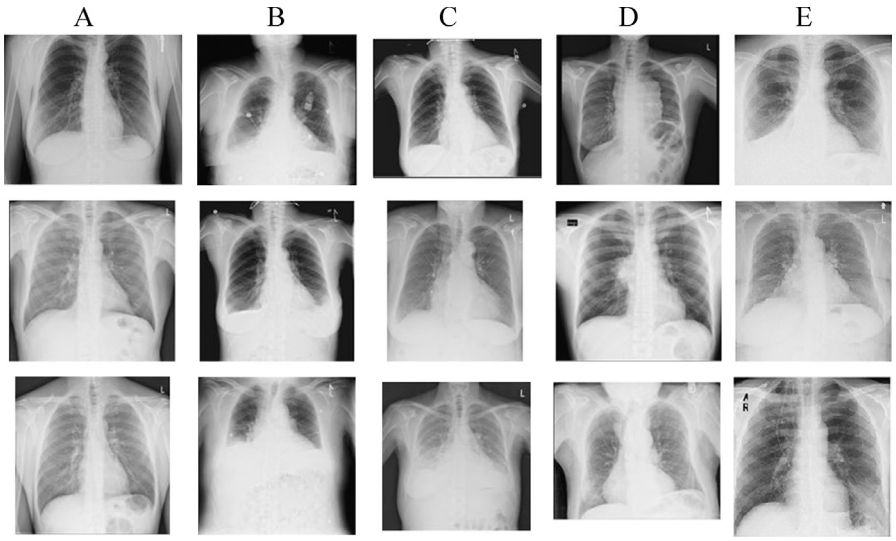


FIGURE 13.1

Chest X-rays categories examples: (A) healthy, (B) left or right pleural effusion, (C) enlarged heart (cardiomegaly), (D) enlarged mediastinum, (E) left or right consolidation.

Source: Diagnostic Imaging Department, Sheba Medical Center, Tel Hashomer, Israel

13.1 INTRODUCTION

With approximately 2 billion procedures per year, radiographs (X-rays) are the most common imaging examination in radiology. Adult chest radiographs are a major part of these procedures. They are the most commonly ordered screening test for pulmonary disorders. Chest radiographs are performed to evaluate the lungs, heart, and thoracic viscera. They are essential for the management of various diseases associated with high mortality, including pneumonia, heart failure, pleurisy, and lung cancer. Diagnosis of the various chest conditions is a difficult task even to human experts, due to the subtlety of the information.

Examples of pathological chest radiographs are shown in Fig. 13.1. Several examples per pathology are shown: *Pleural effusion* is excess fluid that accumulates in the pleural cavity, the fluid-filled space that surrounds the lungs. This excess of fluid can impair breathing by limiting the expansion of the lungs. Enlarged heart, otherwise termed *Cardiomegaly*, produces an abnormally large shadow (cardiac silhouette) on a chest X-ray film. Detection of cardiomegaly is often conducted using the cardiothoracic ratio which is defined as the ratio of the maximal horizontal cardiac diameter and the maximal horizontal thoracic diameter (inner edge of ribs or edge of pleura). The heart is enlarged if the cardiothoracic ratio is greater than 50%. *Mediastinal abnormalities* are important radiological findings. The causes of mediastinal widening can be divided into traumatic and nontraumatic and include vascular abnormality and

FIGURE 13.2

An example of multiple pathologies per patient. The patient is diagnosed with right pleural effusion, enlarged heart, and enlarged mediastinum.

mediastinal mass. *Pulmonary opacification* refers to a visible area of increased attenuation, related to a decrease in the ratio of gas to soft tissue in the lung. This finding often indicates a replacement of air in the alveoli or small airways with dense material (e.g., fluid, blood, cells). Pulmonary opacities may be seen in different kinds of lung diseases such as pneumonia, cancer, and tuberculosis.

Detecting cardiomegaly and pleural effusion in frontal X-rays seems a feasible task since in most cases the heart is either clearly normal in size or clearly abnormally enlarged, and lungs shape and expansion capabilities also stand out from the normal case. Accurate detection of lung consolidation, on the other hand, is a more difficult pathology to detect and characterize correctly. In a clinical setting, it is often the case that *multiple* pathologies are present in a given chest radiograph. An example of such a case is highlighted in Fig. 13.2, illustrating a patient clinically diagnosed with right pleural effusion, enlarged heart, and enlarged mediastinum.

The increasingly growing workload on the radiology staff in radiology departments around the world leads to the reality that radiographs may not be read by the radiologists staff or the read is following a long, sometimes life endangering delay. Therefore, there is an interest in developing automated computer analysis systems to assist the radiologists in the reading task. An automated software which can discriminate between healthy and non-healthy cases is of substantial need, to support initial screening of the examined cases. Moreover, an automated system that shows high sensitivity and specificity in categorizing a set of pathologies can be of great value in real clinical settings. It is expected that such systems can improve accuracy and efficiency in the radiology departments of the Western world, and can be critical to healthcare in third world countries (e.g., China), where no substantial radiology service is available.

A topic of great interest for analysis in chest X-rays is the detection of lung nodules, which are important as a precursor for cancer. Although the target of most research attention, lung nodules are a relatively rare finding in the lungs. The most common findings in chest X-rays include lung infiltrates, catheters, and abnormalities

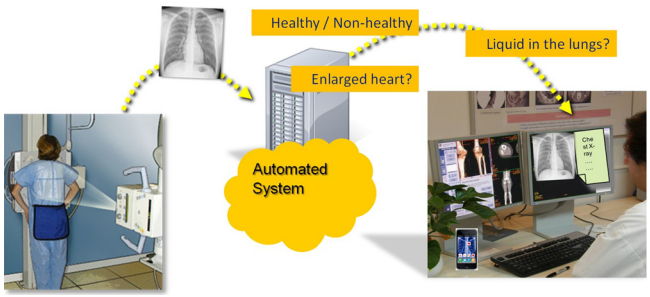


FIGURE 13.3

Automated system for multi-label pathology identification.

of the size or contour of the heart [1]. In our work we focus on the more frequently appearing set of pathologies that need to be written in the radiology report.

Most works found in the literature are *single task* focused. A set of segmentation and landmark localization tools are used to address the specific task at hand. Methods used are often algorithmically and computationally challenging, requiring, for example, segmentation of the lungs, suppression of ribs, and localization of typical lung textures (detection of emphysema [2], diagnosis of interstitial lung diseases [3]).

Very few studies can be found in the literature that focus on chest pathology identification and classification as an *image-level* labeling task (e.g., [4,5]). In [4] healthy versus pathological identification was explored using Local Binary Patterns (LBP). Avni et al. [5] used the Bag-of-Visual-Words (BoVW) model [6] to discriminate between healthy and four pathological cases. The BoVW methodology was proven to lead the ImageClef competitions (http://www.imageclef.org) in categorizing X-rays on the organ level (2008, 2009). The success of the BoVW framework on the organ level led to its extension to pathology level categorization of the chest radiographs [5].

Our objective in the current work is to explore the role of a Deep Learning approach as an automated image-level system for pathology categorization of chest-related diseases and in the screening of healthy versus non-healthy cases. See Fig. 13.3 for illustration.

Deep Learning is a class of machine learning techniques, where many layers of information processing stages in hierarchical supervised architectures are exploited for feature learning and for pattern analysis/classification. The essence of deep learning is to compute hierarchical features or representations of the observed data, where the higher-level features or factors are defined from lower-level ones [7]. Deep (i.e., many-layered) convolutional neural networks (CNNs) for machine object recognition, are advancing the limits of performance in domains as varied as computer vision, speech, and text, and are considered as one of the leading technologies for recognition [8,9]. Recent results indicate that the generic descriptors extracted from CNNs are extremely effective in object recognition and provide better results than

systems relying on carefully engineered representations, such as SIFT or HOG features [10,11].

Deep learning methods are most effective when applied to networks with large number of training data to train the deep neural network. In the computer-vision domain such large image sets exist and enable the training of popular CNNs in many image recognition tasks, such as the large scale visual recognition challenge of ImageNet [12–16]. Other domains exist in which there is less data for training. Recently, works have come out in the general computer-vision literature that use *Transfer Learning* – an efficient way to utilize image representations learned with CNNs on large-scale annotated datasets, defined as the source, to domains in which limited data exists, termed the target domain. Such tools are also entering the medical imaging domain, as can be seen by the emerging set of works in the field. For a general overview see [17].

In this chapter we provide an overview of our exploration into two main issues:

- Can we provide a system to automatically analyze an input radiograph to detect multiple pathologies and to provide *multiple labels* per case? Note that our goal is to screen for healthy or non-healthy, and to categorize all pathologies present in the image, in an efficient and robust framework that can adapt to a real clinical setting.
- Can we utilize the deep learning methodology for this medical task? We assume a scenario in which medical data is limited and no network training can be done. We explore if the general-image trained deep network features, trained on the ImageNet data (source domain), provide a robust representation which we can use to categorize the chest radiograph data (target domain).

We show classification results of six different pathologies, on several hundreds of cases from a real clinical setting. Using the deep learning framework, we show categorization results for all pathologies, while exploring further the deep network features using fusion and selection schemes.

13.2 IMAGE REPRESENTATION SCHEMES WITH CLASSICAL (NON-DEEP) FEATURES

Developing an image recognition solution that is based on a new set of features must be compared against a strong baseline of well established feature extraction techniques. As a benchmark we apply a set of classical descriptors that are known in the literature to perform well in image classification/categorization tasks. These include GIST [18] features, Pyramid Histogram of Oriented Gradients (PHOG) [19], Gabor [20], and Gray-Level Co-occurrence Matrix (GLCM) statistics [21]. A brief overview of these filters is described in Section 13.2.1. We also compare the deep feature representation to the more recent state-of-the-art representation of the Bag-of-Visual-Words (BoVW) model [6]. A brief review of the BoVW is provided in Section 13.2.2.

13.2.1 CLASSICAL FILTERING

The *GIST descriptor* proposed in [18] provides a low dimensional representation of a scene which does not require any form of segmentation. The descriptor focuses on the shape of the scene, on the relationship between the outlines of the surfaces and their properties, and ignores the local objects in the scene and their relationships. It is derived by resizing an image to 128×128 and iterating over different scales (4 scales in our case) where for each scale the image is divided into 8×8 cells. For each cell, orientation (every 45 degrees), color, and intensity histograms are extracted. The descriptor is a concatenation of all histograms, for all scales and cells. The GIST descriptor is similar in spirit to the local SIFT descriptor [24]. It was found to be helpful in scene recognition, e.g., in [22,23].

The *PHOG descriptor*, proposed by Bosch et al. [25], represents an image by its local shape and the spatial layout of the shape. Local shape is captured by the distribution over edge orientations within a region, and spatial layout by tiling the image into regions at multiple resolutions. The descriptor is derived by iterating over different scales (2 scales in our case) where for each scale, the image is first divided into cells and for each cell a histogram of gradient directions or edge orientations is extracted. The descriptor is a concatenation of all histograms, for all scales and cells. PHOG has been successfully applied to object classification in recent years [19,26].

Both GIST and PHOG descriptors are known for capturing the statistical information of a scene, by capturing local shapes based on the distribution of data within the regions of interest, and capturing spatial layout based on the tiling of the image into regions of multiple resolutions.

A closely related descriptor is the Gabor filter set [20]. Gabor descriptors can be considered as edge detectors with adjustable orientations and scales. The descriptors are extracted by creating a Gabor filter bank that comprises a set of Gaussian filters that cover the frequency domain with different radial frequencies and orientations (5 wavelengths and 8 orientations in our case), these filters are convolved with the input image to produce corresponding set of response matrices (40 in our case) of the image. In this work, we use two Gabor-derived representations: A *Gabor-raw descriptor*, which is the result of vectorization of all the response matrices downsampled by a factor of 4, and a *Gabor-statistical-based descriptor* that is generated by combining measurements of mean, standard deviation, energy, and entropy for each response matrix from each scale $(2, 4, 8, 16, 32)$ and orientation $(0, 30, 45, 60, 90, 120, 135, 150)$. Gabor filters have been found to be very effective in texture representation and discrimination [20,27].

The gray-level co-occurrence matrix (GLCM) [21] is a statistical method that considers the spatial relationship of pixels in order to extract texture information. A single GLCM matrix represents how often pairs of pixel with specific values at a specified spatial relationship, occur in an image. Element (i, j) of the GLCM matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j. GLCM matrices with differ-

ent horizontal, vertical, and diagonal offsets are generated. Because GLCM matrices are typically large and sparse, they are often represented by a set of features, known as Haralick features [21], which are extracted to form a unique image descriptor. The GLCM image descriptor is very common in the medical image analysis domain [28,29].

13.2.2 BAG-OF-VISUAL-WORDS MODEL

The BoVW [6] image representation provided in recent years the state-of-the-art image classification method. Adapted from the bag-of-words (BoW) representation of text documents, the BoVW method includes the following steps (Fig. 13.4): (i) Patch extraction – to extract uniform size patches from the image ROIs. Each small patch shows a localized view of the image content. These patches are considered as candidates for basic elements, or visual-words. The patch size needs to be larger than a few pixels across in order to capture higher-level semantics such as edges or corners. At the same time, the patch size should not be too large if it is aimed to serve as a common building block for many images. (ii) Feature description – feature representation involves representing the patches using feature descriptors. Raw image data, normalized raw data, or other descriptors can be used. To reduce the computational complexity of the algorithm and the level of noise, a principal component analysis procedure (PCA) can be applied to this patch collection. The first few components of the PCA, which are the components with the largest eigenvalues, serve as a basis for the information description. (iii) Quantization and clustering – the final step of the bag-of-words model is to convert vector-represented patches into visual words and generate a representative dictionary. A frequently-used method is to perform K-means clustering over the vectors of the initial collection. The vectors are then clustered into groups in the feature space. The resultant cluster centers serve as a vocabulary of visual words. A given image can now be represented by a unique distribution over the generated dictionary of words, which is a representative image histogram.

The BoVW method has been successful in medical classification tasks, such as medical image retrieval [30], retrieval of similar-appearing liver lesions [31], the classification of breast tissue [32], and retrieval of brain tumors in MRI images [33]. Our group participated in the 2009 imageCLEF X-ray categorization competition using the BoVW model, and was ranked first [5]. The model used dense sampling of simple features with spatial content, and a nonlinear kernel-based SVM classifier. Motivated by the success in the competition on the organ-level categorization, we extended the system to pathology-level categorization of chest X-ray data. Additionally, we developed BoVW variants which were used in different tasks such as classification of breast tissue [34], liver lesion classification (dual dictionary model [35]) and multi-phase liver lesion classification (relevant words representation [36]).

As a benchmark for the current study, we use the same implementation described in [5]: We add the patch center coordinates to the feature vector to introduce spatial

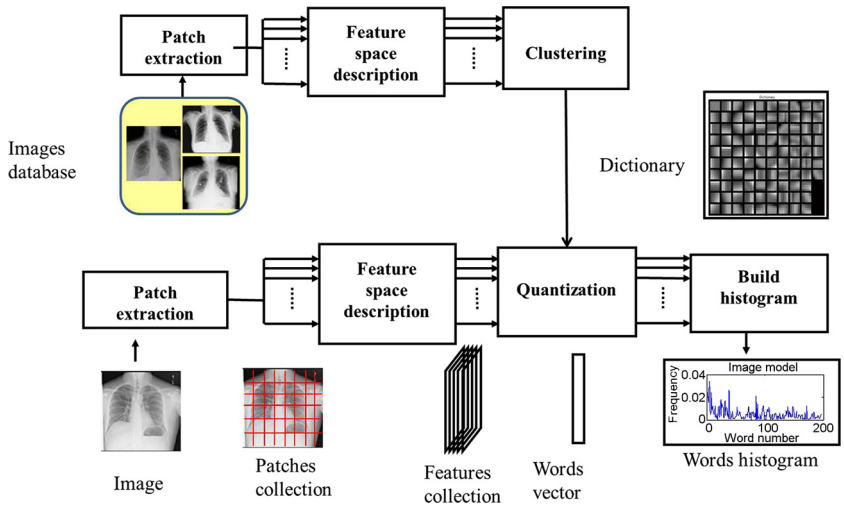


FIGURE 13.4

A schematic illustration of BoVW model.

information to the image representation, without the need to explicitly model the spatial dependency between patches. Empirically, we found that using a 9-dimensional patch representation which comprises 7 PCA components of variance-normalized raw patches along with 2 patch coordinates and using a dictionary size of 1000 words results in best classification performance.

13.3 EXTRACTING DEEP FEATURES FROM A PRE-TRAINED CNN MODEL

CNNs constitute a feed-forward family of deep networks where intermediate layers receive as input the features generated by the former layer and pass their outputs to the next layer. The initial layers of the CNN are locally-connected, alternating convolution and pooling layers, and the final layers of a network are fully-connected. This hierarchy of layers enables different levels of abstraction in the input representation: For visual data, the low levels of abstraction describe the different orientated edges and lines in the image; middle levels describe parts of an object such as corners, angles and surface boundaries, while high layers refer to larger object parts and even complete objects.

In recent years, CNNs have gained considerable interest due to the development of new variants of networks and the advent of efficient parallel solvers optimized for modern GPUs. Advanced computing resources have allowed deeper and more com-

plex convolutional networks to be trained. Numerous CNN architectures that improve the previous state-of-the-art classification/analysis results obtained using shallow representations have been proposed.

A well known model is the Krizhevsky et al. CNN representation [37] which was learned over a subset of images from ImageNet [12], a comprehensive real-life large scale image database (>20M) that is arranged into more than 15K non-medical concepts/categories. A follow up network architecture which was inspired by the Krizhevsky network, was the Zeiler and Fergus [14] CNN model. Zeiler and Fergus network is constructed from five convolutional layers, interleaved with nonlinear and pooling operations, followed by two fully connected classification layers and the fully connected output layer. In their network, Zeiler and Fergus attempted to correct problems in Krizhevsky et al. model without changing its learning framework. Zeiler and Fergus main claim was that the first convolutional layer filters are a mix of extremely high and low frequency information, with little coverage of the mid frequencies. It was also claimed that due to the first convolutional layer large stride, the second layer visualization show aliasing artifacts. Zeiler and Fergus therefore reduced the first layer kernel size and made a decreased stride size, allowing the new architecture to retain much more feature information in the first two layers. It was shown that the architecture slightly outperforms Krizhevsky et al. architecture in the ImageNet classification challenge.

Obtaining datasets in the medical imaging domain that are as comprehensively annotated, such as the ImageNet data, remains a challenge. Transfer learning and fine tuning are key components in the use of deep CNNs in medical imaging applications [17]. In transfer learning, CNN models (supervised) pre-trained from natural image dataset or from a different medical domain are used for a new medical task at hand. In one scheme, a pre-trained CNN is applied to an input image and then the outputs are extracted from layers of the network. The extracted outputs are considered features and are used to train a separate pattern classifier. For instance, in [38–40], pre-trained CNNs were used as a feature generator for chest pathology identification. In [41] integration of CNN-based features with handcrafted features enabled improved performance in a nodule detection system. Fine tuning is relevant when a medium sized dataset does exist for the task at hand. One suggested scheme is to use a pre-trained CNN as initialization of the network, following which further supervised training is conducted, of several (or all) the network layers, using the new data for the task at hand. In this chapter we focus on the use of transfer learning, with no additional network learning.

We show our initial set of results [39] using the Decaf CNN representation which closely follows the network of Krizhevsky et al. [15]. We then proceed to use the VGG-M-2048 pre-trained CNN model [13] which has a similar architecture to the one used by Zeiler and Fergus with the exception of a very minor change in the first two convolutional layers stride size which was introduced to keep the computation time reasonable, along with a reduction in the number of filters in the fourth convolutional layer. A schematic illustration is provided in Fig. 13.5.

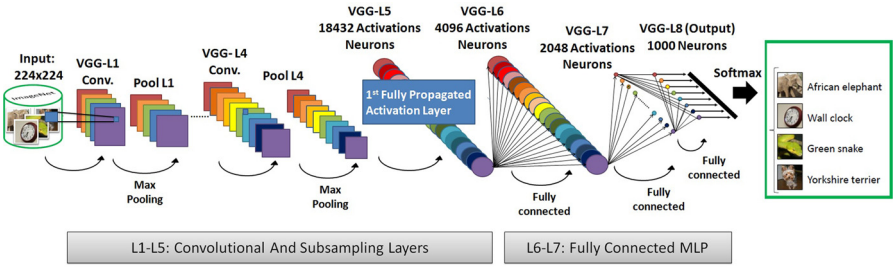


FIGURE 13.5

A schematic illustration of VGG-M-2048 CNN architecture and training process [13].

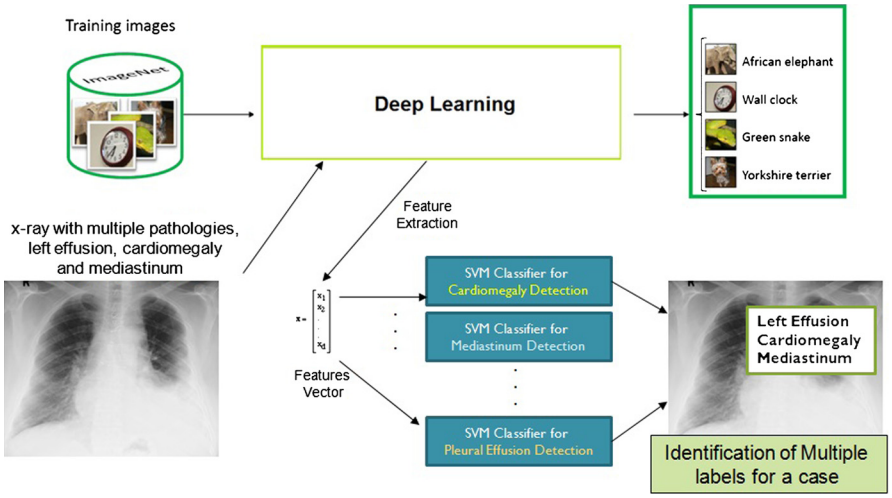


FIGURE 13.6

Identification of multiple X-ray pathologies via transfer learning.

Features from all layers are extracted (excluding the output layer), as the network representation for the identification task. We denote by *VGG-L*1 up to *VGG-L5* the convolutional layers, where *VGG-L5* is the final convolutional layer and is the first set of activations that has been fully propagated through the convolutional layers of the network. *VGG-L6* denotes the first fully-connected layer. We use the notation in [15] to denote the 5th–7th activation layer features of the Decaf network as the network representation.

In the current study, we compare across the different representation schemes, across varying networks. Each selected feature set is input to the SVM classifier, and results are compared. A schematic overview of the full X-ray categorization process using transfer learning is provided in Fig. 13.6.

13.4 EXTENDING THE REPRESENTATION USING FEATURE FUSION AND SELECTION

In further analysis we also investigated variations, such as feature fusion and selection, that can be applied to the features with an attempt to augment results and increase robustness.

Feature selection, a process of selecting a subset of original features according to certain criteria, is an important and frequently used dimensionality reduction technique for data mining. In the past few decades, researchers have developed many feature selection algorithms. These algorithms are designed to serve different purposes, are based on different models, and all have their own advantages and disadvantages. A repository of feature selection algorithms can be found in [42].

In the current work we use the Information Gain feature selection method [43]. Information Gain is a measure of dependence between the feature and the class label. It is one of the most popular feature selection techniques as it is easy to compute and simple to interpret. The Information Gain (IG) of a feature X and the class labels Y is defined as

$$IG(X, Y) = H(X) - H(X|Y), \tag{13.1}$$

the entropy of X and the entropy of X after observing Y, respectively:

$$H(X) = -\sum_{i} P(x_i) \log_2(P(x_i)), \tag{13.2}$$

$$H(X|Y) = -\sum_{j} P(y_j) \sum_{i} P(x_i|y_j) \log_2(P(x_i|yj)). \tag{13.3}$$

A feature with high information gain is more relevant for a given task (with max IG equal to 1). Information gain is evaluated independently for each feature and the features with the top-k values are selected as the relevant features.

The rationale for applying the feature selection scheme is derived from the fact that it is hard to predict in advance which pre-trained CNN layer or layers will be the most powerful feature representation, given a specific task. Indeed, it is known from [14] and [15] that the deeper layers of the network can be used as a powerful image descriptor applicable to other datasets, however, Zeiler and Fergus [14] point out the 7th layer as the most significant layer, Donahue et al. [15] point out the 6th layer, and in our earlier work [39], we have pointed out the 5th layer.

13.5 EXPERIMENTS AND RESULTS

13.5.1 DATA

Our dataset consists of 637 frontal chest X-ray images in DICOM format. The images are of variable size. They are cropped, centered, and contain several artifacts such

as reading directives (e.g., arrows, left/right indicators) and medical equipment, but otherwise were not preprocessed (e.g., equalization, scaling). We have replicated the Intensity channel to support the CNN 3-channel RGB input data expectations. The pertained CNN resizes the images into specific accepted resolution automatically.

The images were collected from the Diagnostic Imaging Department of Sheba Medical Center, Tel Hashomer, Israel. Gold standard was achieved using image interpretation done by two expert radiologists. The radiologists examined all of the images independently and then reached a decision regarding the label of every image. For each image, a positive or negative label was assigned. In cases of disagreement, the image was removed from the dataset.

The images depict 6 chest pathology conditions: Right Pleural Effusion (RPE) – 73 images; Left Pleural Effusion (LPE) – 74 images; Right Consolidation (RCN) – 58 images; Left Consolidation (LCN) – 45 images; Cardiomegaly/Enlarged Heart (Cardio) – 154 images, and Abnormal (Enlarged) Mediastinum (MED) – 145 images. Overall, the dataset contains 325 images with at least one pathological condition.

13.5.2 EXPERIMENTAL SETUP

We performed a binary categorization task for each pathology. Classification was performed using a Support Vector Machine (SVM) classifier with a nonlinear intersection kernel. We investigated the different linear and nonlinear kernels using standard grid-search technique, and empirically selected the nonlinear intersection kernel, which gave slightly better results. For each binary categorization task, cases diagnosed with an examined pathology were labeled as positive cases, and cases diagnosed with the absence of the examined pathology were labeled as negative cases. In the screening scenario, cases diagnosed with the absence of any pathology were labeled as positive cases, whereas cases diagnosed with the presence of at least one of the examined pathologies were labeled as negative cases. We used k-folded cross-validation with $k = 4$. All features used in our work were normalized: each feature across all images had its mean subtracted and was divided by its standard deviation. The algorithm flowchart is shown in Fig. 13.7.

13.5.3 EXPERIMENTAL RESULTS

We start with a comparison across varying representation schemes: from shallow features (PHOG, GLCM, GABOR, and GIST), to medium level representation (BoVW) and deep features of the deep VGG-M-2048 CNN layer features (for layers *VGG-L*1, ..., *VGG-L*7). Table 13.1 presents accuracy results, measured as the area under the ROC curve (AUC).

We also examine the likelihood ratio measurement which is based on the Sensitivity and Specificity metrics [44,45]. In evidence-based medicine, likelihood ratios are used for assessing the value of performing a diagnostic test. Positive likelihood ratio metric is defined as the ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence

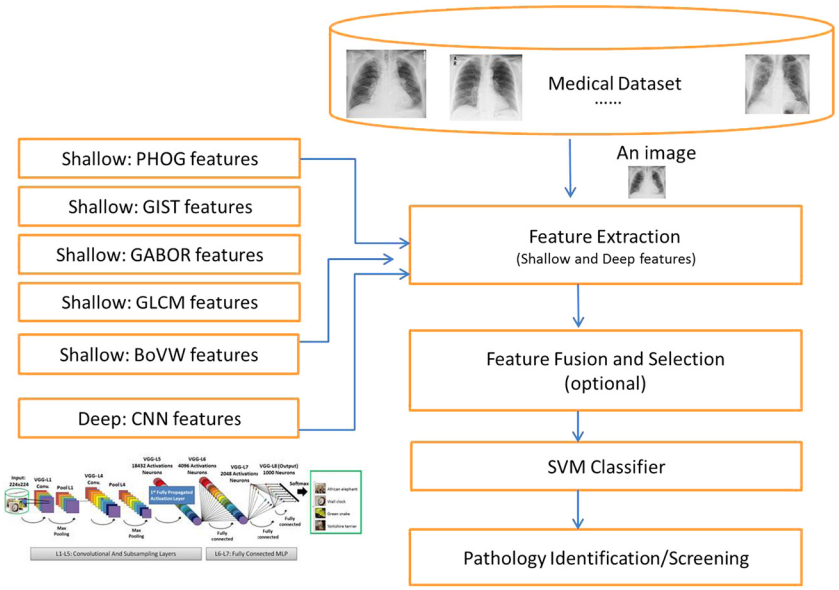


FIGURE 13.7

Algorithm flowchart.

of the disease. This is calculated as

$$LR+ = Sensitivity/(1 - Specificity).$$

The negative ratio metric is defined as the ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease. This is calculated as

$$LR- = (1 - Sensitivity)/Specificity.$$

A likelihood ratio greater than 1 indicates the test result is associated with the disease, a likelihood ratio less than 1 indicates that the result is associated with absence of the disease, and likelihood ratios that lie close to 1 have little practical significance.

From Table 13.1 we observe that for all pathology identification cases, the deep architecture descriptors outperformed the shallow descriptors, with a relatively large performance gap in comparison to statistical based representation such as GLCM and GABOR, and to a lesser extent, in comparison to more sophisticated representation such as BoVW and GIST. CNN intermediate layer features (*VGG-L4*) obtained an average increase of 2–4% AUC over GIST and BoVW features. In terms of $LR+$, *VGG-L4* scored an average of 4.89 against 3.61 and 3.98 for GIST and BoVW features, respectively, reflecting an increase of more than 20%. Similarly, *VGG-L4*

Table 13.1 AUC classification results for classical and network features. The descriptor dimensionality appears in parentheses

Descriptor	RPE	LPE	RCN	LCN	Cardio	MED	Healthy	Avg.
VGG-L4 (86,528)	0.93	0.92	0.77	0.74	0.95	0.85	0.89	0.86
VGG-L5 (18,432)	0.92	0.92	0.77	0.72	0.95	0.86	0.88	0.86
GIST (512)	0.88	0.84	0.82	0.64	0.93	0.80	0.87	0.82
BoVW (1000)	0.90	0.93	0.75	0.70	0.93	0.83	0.87	0.84
PHOG (336)	0.82	0.70	0.71	0.70	0.88	0.71	0.79	0.76
GLCM (440)	0.74	0.68	0.59	0.55	0.80	0.64	0.75	0.68
GABOR-Stat. (160)	0.71	0.69	0.56	0.48	0.82	0.68	0.82	0.68
GABOR-Raw (163,840)	0.89	0.88	0.81	0.65	0.95	0.85	0.89	0.85

Table 13.2 AUC classification results for VGG-M-2048 CNN deep representations. The descriptor dimensionality appears in parentheses

Descriptor	RPE	LPE	RCN	LCN	Cardio	MED	Healthy	Avg.
VGG-L1 (279,936)	0.88	0.88	0.75	0.66	0.95	0.86	0.88	0.84
VGG-L2 (43,264)	0.91	0.91	0.77	0.73	0.96	0.86	0.89	0.86
VGG-L3 (86,528)	0.91	0.92	0.78	0.75	0.95	0.86	0.89	0.87
VGG-L4 (86,528)	0.93	0.92	0.77	0.74	0.95	0.85	0.89	0.86
VGG-L5 (18,432)	0.92	0.92	0.77	0.72	0.95	0.86	0.88	0.86
VGG-L6 (4096)	0.89	0.92	0.79	0.76	0.94	0.86	0.88	0.86
VGG-L7 (2048)	0.89	0.91	0.76	0.79	0.92	0.85	0.87	0.85

scored an average $LR-$ of 0.26, against 0.31 and 0.27 for GIST and BoVW features, respectively, reflecting an improvement as well.

In Table 13.2 we show the AUC classification results for each one of the deep network layers. We can observe that intermediate deep layers, excluding the penultimate layer, provide the strongest representation. In particular, the *VGG-L4* and *VGG-L5* layers of the network. Another observation is that the first VGG layer results in the lowest performance. A strong resemblance in performance can be seen between the *VGG-L1* and the GABOR raw representation performance, for the different tasks. The first convolutional layer applies many filters to its natural image input. This extracts different features of the input which are normally tuned to edges of different orientations, frequency, and color. Due to the fact that our data comprised a replicated intensity channel, the features that are extracted by the network resemble features that are extracted using Gabor filters. This is a behavior that was exhibited in many modern deep neural networks: when trained on images, they all tend to learn first-layer features that resemble either Gabor filters or color blobs [46]. This also provides a biologically-inspired explanation that links the CNN to the human visual system.

We next compared two well known network schemes: the Decaf network and the VGG. In Table 13.3 we present results of layers 5 through 7 of the Decaf CNN (*Decaf-L5*, *Decaf-L6*, *Decaf-L7*). These results are slightly worse than when using

Table 13.3 AUC classification results: comparing between deep networks. The descriptor dimensionality appears in parentheses

Descriptor	RPE	LPE	RCN	LCN	Cardio	MED	Healthy	Avg.
DECAF-L5 (9216)	0.93	0.90	0.77	0.77	0.94	0.85	0.89	0.86
DECAF-L6 (4096)	0.92	0.89	0.76	0.80	0.93	0.86	0.89	0.86
DECAF-L7 (4096)	0.90	0.87	0.73	0.79	0.91	0.84	0.87	0.84

the VGG-M-2048 CNN features: an average $LR+$ of 4.89 was obtained for *VGG-L4* with 4.60 for *DECAF-L5* features. This reflects an increase of approximately 6% with the same $LR-$ measurement of 0.26; At the same average AUC of 0.86, *VGG* intermediate layers perform better in the identification of the Left Pleural Effusion pathology than *Decaf*, with a 2% AUC increase. As the VGG network consistently held the higher scores, we hereon focus on results using the VGG network.

Examples of ROC curves are shown in Fig. 13.8 for the right and left effusion, cardiomegaly, and abnormal mediastinum pathology identification tasks. In all cases the network features provide either the strongest or a match to the strongest representation.

13.5.3.1 Feature Selection Analysis

So far we observed the strength of deep *VGG* layers features for the various identification tasks. We also note fluctuations between the different layers, with an average AUC gap of 3% (ranging from 1 to 6%) across all pathologies. Although *VGG-L4* and *VGG-L5* present strong results, no single layer presents the best score consistently across all pathologies. We next wish to explore feature selection as a means for increasing robustness to the labeling tasks. We use the following steps: (i) *VGG* features from all layers, excluding the first convolutional layer, are fused together, in what is known as fusion of features. The features across all layers (L2–L7) are concatenated into an extended feature-vector. (ii) Feature selection is performed on the joint feature vector in order to remove redundant features. And (iii) the selected set of features are input to the SVM classifier. Table 13.4 shows the AUC classification results of using the top 2500 extracted features, as well as when using the optimally extracted feature set of up to 25,000 extracted features. The optimal number of features, out of the entire augmented 240K+ deep *VGG* feature set, were selected using a standard grid-search technique.

In Table 13.4 we see that using a feature selection scheme can increase performance slightly, by an average AUC gain of 2%. For most of the pathologies, the fixed feature selection representation achieves almost the maximum AUC result, ranking either as the first or second score. In terms of likelihood measurements, the average $LR+$ of the fixed feature selection method is 5.00, as opposed to 4.89 and 4.78 for *VGG-L4* and *VGG-L5* features, respectively. The average $LR-$ of the fixed feature selection method is 0.23, as compared with 0.26 and 0.25. This indicates an improvement over *VGG-L4* and *VGG-L5* as well.

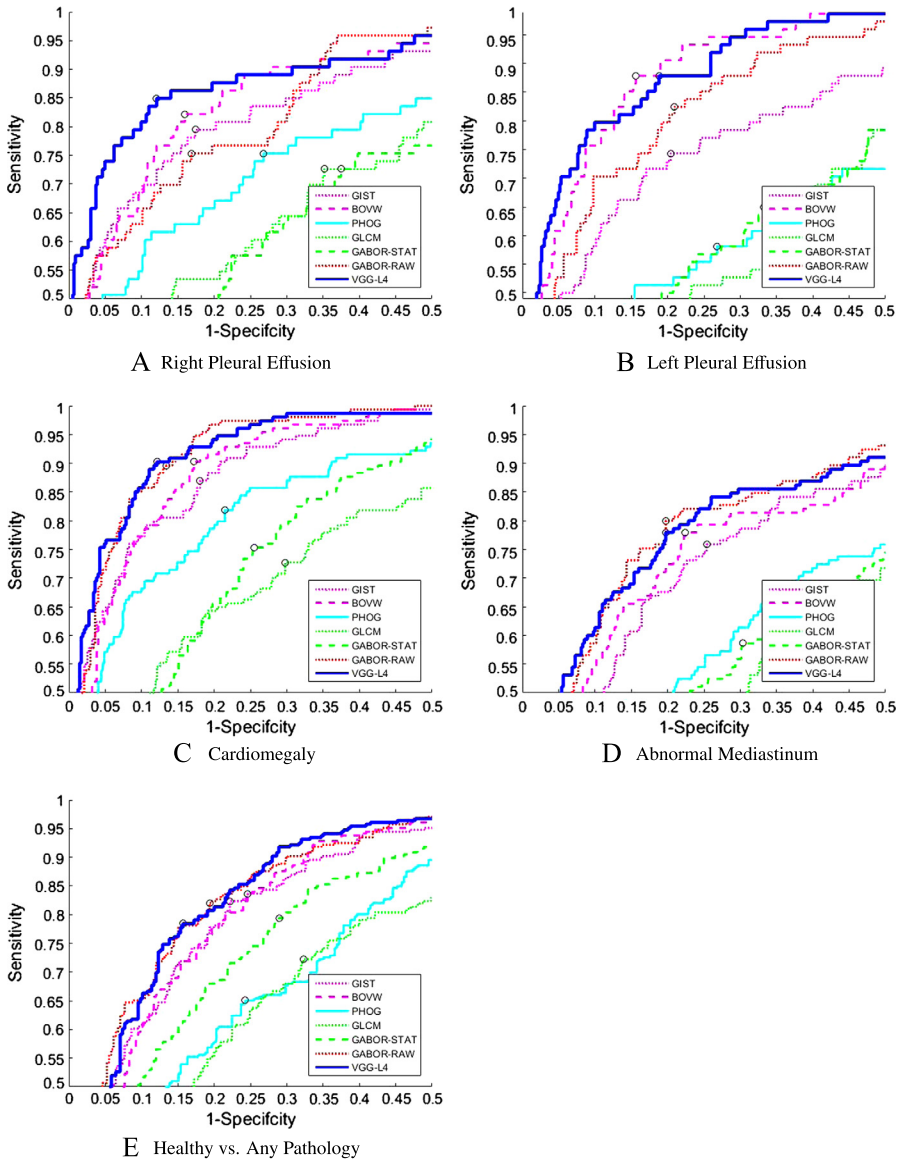


FIGURE 13.8

ROCs of different examined pathologies.

Fig. 13.9 shows a comparative ROC curve analysis using a selected optimal feature set vs the individual layer of *VGG-L4*, for several pathologies. We note that in all examined cases, for most points on the curve, the feature selection curve is on top.

Table 13.4 VGG-M-2048 CNN top ranked deep features representations: AUC metric classification performance. The descriptor dimensionality appears in parentheses

Descriptor	RPE	LPE	RCN	LCN	Cardio	MED	Healthy	Avg.
Selection (2500)	0.92	0.94	0.80	0.78	0.95	0.87	0.89	0.88
Selection (optimal)	0.93 (22,000)	0.94 (1500)	0.83 (400)	0.81 (300)	0.96 (24,500)	0.88 (8000)	0.89 (24,000)	0.89

We obtain similar results, excluding the cardiomegaly case, using the fixed (2500) number of features.

Fig. 13.10 presents a case study analysis for the *VGG-L4* feature representation using the SVM classifier. Four subjects are shown. For each one we compare the radiologist ground truth prediction with the classifier normalized score. The 4 examples shown represent a case of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), respectively. As this figure shows, it is often the case that images that are evident to the human observer are classified with high probability to the correct label, whereas cases that are difficult (noisy) for the human observer are the ones for which the automated system may misclassify as well.

13.6 CONCLUSION

In this chapter we showed a collection of results that demonstrate the robustness and strengths of using the CNN representation for medical categorization tasks. We achieve strong results for medical image classification of various pathologies, which surpass relatively recent state-of-the-art results using BoVW. We explore features for the chest pathology categorization task. We use both classical as well as CNN-based features. We show results for categorization based on each feature set independently, as well as the result of using a feature selection stage prior to the categorization task.

How does the deep learning approach compare with the more classical approaches for medical image analysis? A classical approach for medical processing involves preprocessing of the image, segmentation, and landmark extraction, prior to feature extraction and an optional classification stage. As an example, we refer to a recently published work on the detection of pleural effusion [47]. The solution proposed is based on the use of expert-knowledge for the specific pathology. Specifically, segmentation and landmark localization are used based on which localized features are extracted and later classified. A numerical comparison is difficult as different datasets are used. Still, it is interesting to note that for a similar size dataset, with overall similar clinical conditions, the more classical approach in [47] achieves AUC of 0.9 and 0.85 for the correct categorization of RPE and LPE, and our presented deep learning

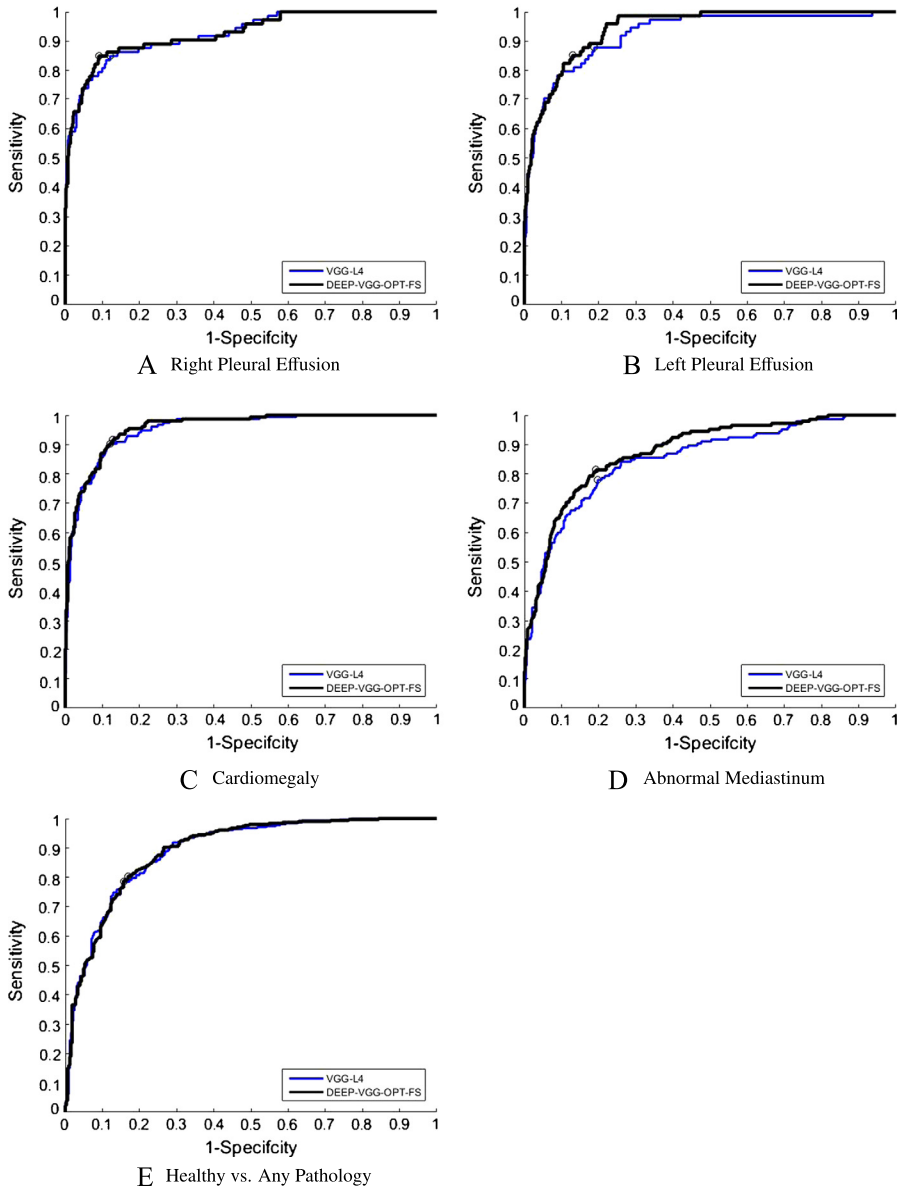
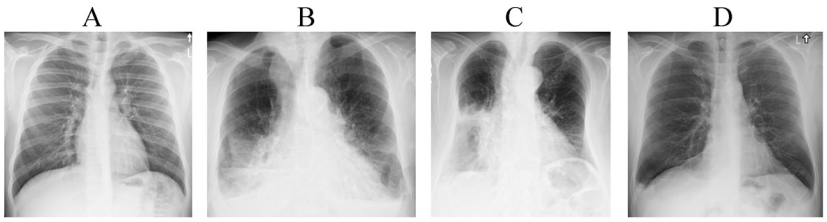


FIGURE 13.9

ROCs of different examined pathologies for the optimal number of selected features.



	A	B	C	D
Normalized Classifier Score ([0,1] range)	0	1	0.935	0.002
Gold Standards	0	1	0	1
	TN	TP	FP	FN

FIGURE 13.10

Right pleural effusion case study analysis using *VGG-L4* feature representation.

based representation scheme achieves similar results of AUC of 0.92 and 0.94, respectively. In the method presented here, the pleural-effusion pathology is just one of several tasks which are *simultaneously* categorized. Also, the use of segmentation and localization are computationally intensive and prone to error, whereas in the framework presented herein, the full image is input to the system with no segmentation necessary.

We conclude that the most informative feature set consists of a selection of features from the CNN layers. Using this selected set of features gives higher AUC values across all pathologies as well as for screening (healthy vs. pathological). Intuitively one could argue that the learned weights which constitute the deep feature layers are optimized to the images of the CNN training dataset and the task it is trained for, thus, one could imagine the optimal representation for each problem lies at an intermediate layer of the CNN but without knowing which layer in advance. Feature selection algorithms can assist us in this problem by picking the most significant deep features from the different layers, in an automated and robust process, while preserving and augmenting the classification performance. The presented approach is general and can be applied to many additional medical classification tasks.

ACKNOWLEDGEMENTS

Part of this work was funded by the Ministry of Industry, Science and Development (Kamin program). Partial support was also given by INTEL Collaborative Research Institute for Computational Intelligence (ICRI-CI).

REFERENCES

1. B. van Ginneken, L. Hogeweg, M. Prokop, Computer-aided diagnosis in chest radiography: beyond nodules, Eur. J. Radiol. 72 (2) (2009) 226–230.
2. G. Coppini, M. Miniati, M. Paterni, S. Monti, E.M. Ferdeghini, Computer-aided diagnosis of emphysema in COPD patients: neural-network-based analysis of lung shape in digital chest radiographs, Med. Eng. Phys. 29 (2) (2007) 76–86.
3. S. Katsuragawa, K. Doi, Computer-aided diagnosis in chest radiography, Comput. Med. Imaging Graph. 31 (4) (2007) 212–223.
4. J.M. Carrillo de Gea, G. Garcia-Mateos, Detection of normality/pathology on chest radiographs using LBP, in: Proceedings of BioInformatics, Valencia, Spain, January 20–23, 2010, pp. 167–172.
5. U. Avni, H. Greenspan, E. Konen, M. Sharon, J. Goldberger, X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words, IEEE Trans. Med. Imaging 30 (3) (2011) 733–746.
6. G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints. Workshop on statistical learning in computer vision, in: ECCV, vol. 1, Prague, 2004, pp. 1–2.
7. Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
8. J. Dean, Large scale distributed deep networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1223–1231.
9. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556, 2014.
10. A.S. Razavian, A. Hossein, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2014, pp. 512–519.
11. M. Oquab, et al., Learning and transferring mid-level image representations using convolutional neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1717–1724.
12. J. Deng, et al., ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
13. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, arXiv:1405.3531, 2014.
14. M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision – ECCV 2014, Springer, 2014, pp. 818–833.
15. J. Donahue, et al., Decaf: a deep convolutional activation feature for generic visual recognition, arXiv:1310.1531, 2013.
16. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localization and detection using convolutional networks, arXiv:1312.6229, 2013.
17. H. Greenspan, B. van-Ginneken, R.M. Summers, Deep learning in medical imaging: overview and future promise of an exciting new technique, IEEE Trans. Med. Imaging 35 (5) (2016) 1153–1159.
18. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
19. S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2 (2006) 2169–2178.

20. I. Fogel, D. Sagi, Gabor filters as texture discriminator, Biol. Cybern. 61 (2) (1989) 103–113.
21. R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 6 (10) (1973) 610–621.
22. M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, C. Schmid, Evaluation of GIST descriptors for web-scale image search, in: Proceedings of the ACM International Conference on Image and Video Retrieval, Springer, 2009, pp. 19:1–19:8.
23. R. Raguram, C. Wu, J.M. Frahm, S. Lazebnik, Modeling and recognition of landmark image collections using iconic scene graphs, Int. J. Comput. Vis. 95 (3) (2011) 213–239.
24. D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
25. A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Springer, 2007, pp. 401–408.
26. Y. Bai, L. Guo, L. Jin, Q. Huang, A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition, in: Proceedings of the 16th IEEE International Conference on Image Processing, Springer, 2009, pp. 3269–3272.
27. B. Ioan, G. Alexandru, Directional features for automatic tumor classification of mammogram images, Biomed. Signal Process. Control 6 (4) (2011) 370–378.
28. A. Karahaliou, S. Skiadopoulos, I. Boniatis, P. Sakellaropoulos, E. Likaki, G. Panayiotakis, L. Costaridou, Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis, Br. J. Radiol. 80 (956) (2007) 648–656.
29. Y. Chi, J. Zhou, S.K. Venkatesh, Q. Tian, J. Liu, Content-based image retrieval of multiphase CT images for focal liver lesion characterization, Med. Phys. 40 (10) (2013).
30. Sebastian Haas, René Donner, Andreas Burner, Markus Holzer, Georg Langs, Superpixel-based interest points for effective bags of visual words medical image retrieval, in: Proceedings of the Second MICCAI International Conference on Medical Content-Based Retrieval for Clinical Decision Support, 2012, pp. 58–68.
31. W. Yang, Z. Lu, M. Yu, M. Huang, Q. Feng, W. Chen, Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single- and multiphase contrast-enhanced CT images, J. Digit. Imaging 25 (6) (2012) 708–719.
32. J. Wang, Y. Li, Y. Zhang, H. Xie, C. Wang, Bag-of-features based classification of breast parenchymal tissue in the mammogram via jointly selecting and weighting visual words, in: Proceedings of the International Conference on Image and Graphics, Hefei, P.R. China, 2011, pp. 622–627.
33. M. Huang, W. Yang, M. Yu, Q. Feng, W. Chen, Retrieval of brain tumors with region-specific bag-of-visual-words representations in contrast-enhanced MRI images, Comput. Math. Methods Med. (2012) 280538.
34. I. Diamant, J. Goldberger, H. Greenspan, Visual words based approach for tissue classification in mammograms, Proc. SPIE Med. Imaging 8670 (2013) 867021.
35. I. Diamant, A. Hoogi, C.F. Beaulieu, M. Safdari, E. Klang, M. Amitai, H. Greenspan, D.L. Rubin, Improved patch based automated liver lesion classification by separate analysis of the interior and boundary regions, J. Biomed. Health Inf. (2015), http://dx.doi.org/10.1109/JBHI.2015.2478255.
36. I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, Multi-phase liver lesions classification using relevant visual words based on mutual information, in: IEEE Int. Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 2015, pp. 407–410.

37. A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
38. Y. Bar, I. Diamant, L. Wolf, H. Greenspan, Deep learning with non-medical training used for chest pathology identification, Proc. SPIE Med. Imaging (2015) 94140V.
39. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, H. Greenspan, Chest pathology detection using deep learning with non-medical training, in: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI), 2015.
40. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, H. Greenspan, Chest pathology identification using deep feature selection with non-medical training, Comput. Methods Biomech. Biomed. Eng. (2016) 1–5.
41. B. van Ginneken, A.A. Setio, C. Jacobs, F. Ciompi, Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI), 2015, pp. 286–289.
42. Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu, Advancing feature selection research, ASU Feature Selection Repository, 2010, pp. 1–28.
43. T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, 2012.
44. J.R. Thornbury, D.G. Fryback, W. Edwards, Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information 1, Radiology 114 (3) (1975) 561–565.
45. S.G. Pauker, J.P. Kassirer, Therapeutic decision making: a cost–benefit analysis, N. Engl. J. Med. 293 (5) (1975) 229–234.
46. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems, 2014, pp. 3320–3328.
47. P. Maduskar, R. Philipsen, J. Melendez, E. Scholten, D. Chanda, H. Ayles, C.I. Sánchez, B. van Ginneken, Automatic detection of pleural effusion in chest radiographs, Med. Image Anal. 28 (2016) 22–32.