A generalized framework for medical image classification and recognition

M. Abedini
N. C. F. Codella
J. H. Connell
R. Garnavi
M. Merler
S. Pankanti
J. R. Smith
T. Syeda-Mahmood

In this work, we study the performance of a two-stage ensemble visual machine learning framework for classification of medical images. In the first stage, models are built for subsets of features and data, and in the second stage, models are combined. We demonstrate the performance of this framework in four contexts: 1) The public ImageCLEF (Cross Language Evaluation Forum) 2013 medical modality recognition benchmark, 2) echocardiography view and mode recognition, 3) dermatology disease recognition across two datasets, and 4) a broad medical image dataset, merged from multiple data sources into a collection of 158 categories covering both general and specific medical concepts—including modalities, body regions, views, and disease states. In the first context, the presented system achieves state-of-art performance of 82.2% multiclass accuracy. In the second context, the system attains 90.48% multiclass accuracy. In the third, state-of-art performance of 90% specificity and 90% sensitivity is obtained on a small standardized dataset of 200 images using a leave-one-out strategy. For a larger dataset of 2,761 images, 95% specificity and 98% sensitivity is obtained on a 20% held-out test set. Finally, in the fourth context, the system achieves sensitivity and specificity of 94.7% and 98.4%, respectively, demonstrating the ability to generalize over domains.

Introduction

Medical image data has been growing by 20% to 40% every year [1], whereas the number of physicians per capita in the United States has remained relatively flat since the 1990s [2]. This trend makes automatic classification and categorization of medical images important so that clinicians are better able to handle the increasing workload demands placed on them. The primary challenge for medical image recognition is its broad domain. Medical images vary by type [e.g., illustration, radiological, electrocardiogram (ECG), or dermatological photograph], modality [e.g., x-ray, magnetic resonance imaging (MRI), ultrasound, or dermoscopy], body region [e.g., brain, heart, chest, or arms], view [e.g., anterior-posterior or superior-inferior], and disease [e.g., infarct, melanoma, or healthy], giving rise to potentially millions of categories. In order for a modeling system to cover such a diverse range of categories, it must be arbitrarily scalable along three orthogonal dimensions: the number of categories the system can model, the amount of data the system can handle (in terms of images), and number of algorithmic approaches the system can utilize. The third dimension is especially important, as no one approach can perform well across all domains in this problem space: an effective system must involve a number of approaches, each of which can be selected based on which best recognizes or models a given category.

In order for research and development to take place in the space of medical image retrieval, datasets must be constructed from which experiments can be performed. Such datasets must contain both image data, as well as labeling structures that are expressive enough to comprehensively capture the diagnostic and categorical information contained within. In addition, these datasets must be large enough to adequately cover the visual variety of the various domains they represent. Some recent public datasets that seek to address these requirements have been Image Retrieval in Medical Applications (IRMA) [3] and the

Digital Object Identifier: 10.1147/JRD.2015.2390017

ImageCLEF (Cross Language Evaluation Forum) Medical Task [4].

IRMA is a comprehensive hierarchical coding structure covering medical image modality, body region, and view. This data structure has been initially populated with data limited to projection radiography (x-ray). Several approaches have been implemented and tested with respect to the dataset. One of the more effective methods has involved the use of classifiers trained from Bags-of-Visual-Word histogram features aggregated over images patches [5]. Other recent work with this dataset, including SPIRS-IRMA [6, 7] (Spine Pathology and Image Retrieval System—Image Retrieval in Medical Applications), has attempted to fuse image level modeling approaches with more localized and specialized methods, such as segmentation approaches, that are capable of quantifying more subtle disease states. While these methods, which are based on common techniques in computer vision and multimedia, have performed well on this initial data population of IRMA, it is unclear from the published experiments how they would scale, in terms of recognition performance, to other domains within medicine.

The ImageCLEF Medical Task is another well-known medical imaging benchmark that has established various challenges over time. The IRMA dataset was featured in this task over a number of years (2005 to 2009). More recently, ImageCLEF has begun to focus predominantly on medical image modality recognition, using a subset of data from a large corpus of PubMed* articles. The image data covers a larger breadth of modalities than IRMA, which was limited to projection x-ray imaging; however, the hierarchy is shallow, not discriminating between body regions, views, or diseased states. A number of algorithmic approaches were evaluated with respect to this dataset, with the best ones achieving in the range of 75% to 80% multiclass accuracy [4, 8].

Other studies have also been spawned in more specialized areas of medical image retrieval and analysis. These include echocardiography view recognition [9, 10] and dermatology image recognition of melanoma.

The determination of echocardiography view and mode is an essential step in automatic cardiac echo image analysis. In an echocardiogram exam, the position and angulation of a 2D ultrasound probe changes the views, and consequently may change the most appropriate analytic algorithm to be applied. Many recent studies have been performed in this area. Ebadollahi et al. [11] proposed a chamber detector based on a generic cardiac chamber template. They used a Markov Random Field (MRF) to locate the chambers and a multi-class Support Vector Machine (SVM) classifier to predict the chamber view. In this approach, the end-diastolic (ED) frame, where the heart is most dilated, is a key clue for identifying the structure and location of chambers. Zhou et al. [12] introduced a similar approach by using ED keyframes and boosted weak classifiers of Haar-like local rectangle features for identifying heart structure and view classification. Otey et al. [13] implemented a hierarchical classification model to identify view type in the upper level, and then view classification in the lower level. They fed well-recognized low-level features, such as gradient, peak, raw pixels, and other statistical features, extracted from images to a Logistic Decision Tree at both levels. Park et al. [14] extended the work by using an MLBoost learning algorithm. Beymer et al. [9, 15] proposed incorporating motion information, by using Active Shape Models (ASMs), for view classification, extracting the shape and texture information, and then tracking these across different frames to derive motion information. Kumar et al. [16] extended this approach by using vocabulary-based PMK (Pyramid Match Kernel) and multiclass. González et al. [17] used a multilayer neural network for view prediction. In recent work, Wu et al. [18] built SVM classifiers trained on extracted Gist low-level features, as Gist provides a global description of the image. Agarwal et al. [19] proposed using SVMs modeled over Histogram of Oriented Gradients (HOG) as a low-level feature. HOG extracts structural information using a set of local histograms.

In dermatology, image analysis is an important problem, as skin cancer is the most common form of cancer in the United States [20]. Every year, over 3.5 million incidences are reported in 2 million people. Every 57 minutes, one person dies of melanoma. Survival rate is 98% with early detection, and 62% when disease reaches lymphatics. Diagnosis of disease is still a very subjective process, varying within clinicians, across clinicians, and between institutions. Quality of care can quickly degrade with lack of experience. Even when expertise is available, accuracy of dermatology experts is estimated between 75% and 84% [21], which leads both to missed diagnoses as well as to unnecessary and potentially disfiguring biopsies. As a result, a multitude of recent work has strived to develop objective image analysis software to aid in the diagnosis of melanoma.

One such work by Barata et al. [22] has endeavored to explore a subset of modeling algorithms toward melanoma detection. In this work, Barata et al. break down the approach into segmentation, feature extraction, and modeling. Within each, learning parameters are varied, and the impacts on performance are assessed over a dataset of 176 images (25 melanomas and 151 nevi). In the segmentation phase, two types are attempted: a whole-lesion segmentation and an interior-lesion segmentation with a border segmentation. The latter mimics the commonly used pyramid or other multi-granularity approaches of the computer vision field—features are extracted from each region independently, and concatenated via early feature fusion. In the feature extraction phase, two relatively simple features are utilized: edge and color histograms. These are either extracted at a global scale, or locally pooled into a histogram using the common bag-of-visual-words approach. In the modeling

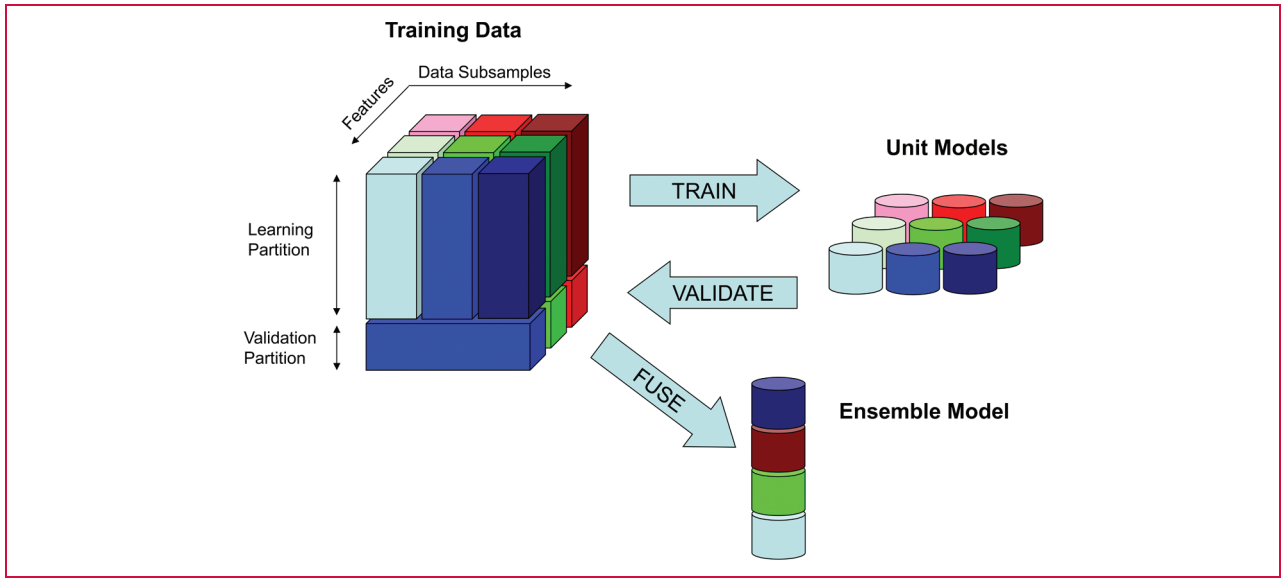

Figure 1

Overview of the IMARS training visual pipeline. Training data is partitioned into Learning and Validation sets. Unit Models are trained on the Learning partition. Fusion of unit models is optimized on the Validation partition.

phase, three variants were studied: kNN (k-Nearest Neighbors), SVM, and AdaBoost. For each, parameters are also varied to optimize the technique. In general, color histograms were found to outperform edge histograms, local pooling was found to improve over global methods, and multi-region segmentation was found to improve over single region. Top performance measures were in the realm of 96% and 80% for sensitivity and specificity. While the achieved performance in this work is quite high, the results are measured using a leave-one-out procedure on a very limited dataset of under 200 images; therefore, whether the conclusions of the work generalize to larger datasets remain unknown.

Garnavi et al. [23] developed a computer-aided diagnostic system for melanoma; the segmentation method [24] involved a hybrid of global thresholding to detect an initial boundary of the lesion, and then application of an adaptive histogram thresholding on optimized color channels of X (from the CIE XYZ color space) to refine the border. The system applied a combination of texture and border-based features, and utilized Gain Ratio to identify optimal features to be used in the classification of melanoma lesions. This approach achieved a significant reduction in the dimension of the feature space (by a factor of 1,542), while increasing the accuracy by 12% and decreasing the computational time by a factor of 50. Applying a random forest classifier on a set of 289 dermoscopy images (114 malignant, 175 benign) partitioned into train, validation, and test image sets, the system achieved an accuracy of 91.26% and an area-under-curve value of 0.937, using 23 optimal features. Experiments demonstrated higher contribution of texture features than border-based features in the optimized feature set.

Commercial products for melanoma recognition have also been developed and subjected to U.S. Food and Drug Administration (FDA) clinical trial. One such product is MelaFind*, which has been studied on a dataset of 1,632 images (175 melanoma), which achieved high sensitivity (98.3%) but low specificity (10.8%), making it difficult to adopt in practice [25].

In this work, we present a visual modeling architecture that is sufficiently flexible and scalable to cover a wide spectrum of domains in classification of medical images. In "Visual modeling approach," we describe the algorithms and implementation, which is based on a two-stage ensemble approach that is implemented in the Hadoop* Map-Reduce parallelization framework for arbitrary scalability. In "Datasets," we describe the datasets in which we evaluate our architecture, covering the following four medical imaging domains: 1) the ImageCLEF2013 medical image modality classification benchmark, a space in which our algorithm is directly comparable to other algorithms designed for this task, 2) a specialized task of echocardiography view and mode recognition, 3) a specialized task of melanoma recognition in two datasets, including the previously mentioned dataset [22], as well as a more recent dataset of over 2,000 images, and 4) a broad medical image category recognition dataset, where we merged multiple datasets into

a collection of 158 categories covering both general and specific medical concepts including modalities, body regions, views, and disease states. In the section "Experimental results," we describe our experiments and results for each of these datasets, and the paper ends with a conclusion.

Visual modeling approach

We study the efficacy of a variant of an ensemble modeling approach [26, 27]. Specifically, we implement several improvements to the IMARS visual learning framework [28, 29]. IMARS is a two-stage ensemble learning pipeline, whereby training data is partitioned into a "Learning" and "Validation" partition (**Figure 1**). A variety of low-level features are extracted (and normalized) over the training data, including color histogram, edge histogram, Gist, color correlogram, and LBPs (Local Binary Patterns), among other global and local descriptors [30–35]. Some of the features used have been reported in prior literature related to medical image modality classification [3]. Each feature is extracted over a variety of spatial granularities, such as global (entire image), horizontal parts (three equally sized horizontal segments), and layout (four quarters, and the image center). Then, for each category, unit models are trained on subsets of data and single feature types from the Learning partition. These subsets are referred to as "bags." For each unit model training task, the system may use a variety of machine learning algorithms and parameters, specified by the user, and optimize the selection based on cross-fold validation. Once unit models are trained, they are then input to a forward model selection learning process on the Validation set. Forward model selection will automatically determine an optimal ensemble of unit models (data and features), relieving the user of any guess-work into what features should be used for modeling. This is done by first initiating the ensemble model with the single unit model that achieves best performance on the Validation dataset. Subsequently, a search is performed to find the unit model that, when combined with the existing ensemble, boosts performance the most. This process is continued until performance saturates.

For large-scale training and scoring, we used the Hadoop Map-Reduce implementation of IMARS for large-scale ensemble classifier learning [28, 29]. In this method, unit models are learned in the Hadoop Map stage, where each task is independent, and ensembles are optimized in the Hadoop Reduce stage, where independent tasks are aggregated. Classifier scoring happens in a likewise fashion: unit models are scored against image features in a Map stage, and ensembles of unit model outputs are aggregated in the Reduce stage. A physical cluster of approximately 800 CPU cores (~700 used for data processing and ~100 reserved for OS tasks), 3.1 TB of total system memory, and 70 TB of hard disk storage is used for experiments.

For the purposes of this work, several improvements to the IMARS visual learning pipeline have been implemented.

Table 1 Categories of the ImageCLEF 2013 benchmark.

Acronym	*Category*
COMP	Compound or multipane images (one category)
D3DR	3D reconstructions (one category)
DMEL	Electron microscopy
DMFL	Fluorescence microscopy
DMLI	Light microscopy
DMTR	Transmission microscopy
DRAN	Angiography
DRCO	Combined modalities in one image
DRCT	Computerized tomography
DRMR	Magnetic resonance
DRPE	PET
DRUS	Ultrasound
DRXR	X-ray
DSEC	Electrocardiography
DSEE	Electroencephalography
DSEM	Electromyography
DVDM	Dermatology
DVEN	Endoscopy
DVOR	Other organs
GCHE	Chemical structure
GFIG	Statistical figures
GFLO	Flowcharts
GGEL	Chromatography
GGEN	Gene sequence
GHDR	Hand-drawn sketches
GMAT	Mathematics
GNCP	Non-clinical photos
GPLI	Program listing
GSCR	Screenshots
GSYS	System overviews
GTAB	Tables and forms

These include low-level features, modeling algorithms, score normalization, and synthetic minority oversampling techniques.

The first group of enhancements involves additional sets of spatial granularities in which features are extracted. These include "pyramid" and "pyramid23." "Pyramid" granularity is a spatial pyramid with global scope as the first level (1×1 image grid), followed by a 2×2 image grid as the second level (which increases feature dimensionality by a factor of 5). "Pyramid23" uses global (1×1) as first level, 2×2 image grid as second level, and 3×3 as the third level (which increases feature dimensionality by a factor of 14).

The second group of enhancements involves additional implemented visual features. In addition to those previously involved in the IMARS framework, we add variations of

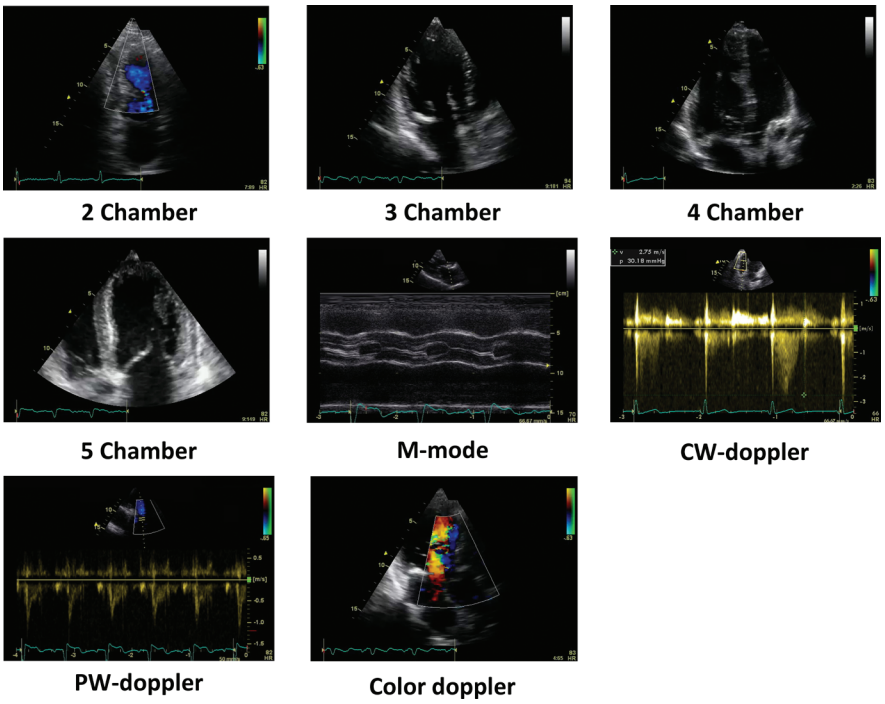

Figure 2

Examples of echocardiography view and mode. Two-dimensional (2D) mode two- to four-chamber views depict varying numbers of cardiac anatomical chambers. The five-chamber view includes visualization of the aortic outflow tract. M-mode refers to "Motion Mode," where movement of anatomical surfaces is visualized. Continuous wave Doppler (CW-Doppler) samples a line through the body, whereas pulsed wave localizes sampling within a small volume. Color Doppler overlays Doppler information on a 2D mode image.

Multiscale and Multi-Color Channel LBPs and Spatially Invariant Feature Transforms (SIFT), as well as Fourier Polar Pyramids. Color LBP [30, 31] is an extension of the common grayscale LBP, whereby LBP descriptors are extracted across five color channels (Red, Green, Blue, Saturation, and Hue), with one histogram per color channel. For a 59-bin (58 uniform and 1 non-uniform) LBP histogram, this results in $59 \times 4 = 236$ total bins. A full 256 bin variant is also extracted. Multiscale LBP (which can be implemented in conjunction with Multi-color) is implemented by extracting LBP descriptors over various image sizes, and aggregating the descriptors into the same histogram bins, weighted by the inverse of the image size.

SIFT constitutes descriptors extracted around Harris Laplace interest points. Each keypoint is described with a 128-dimensional vector containing oriented gradients. We obtain a visual word dictionary of size 1,000 by running K-means clustering on a random sample of approximately 300,000 interest point features from the ImageCLEF 2013 PubMed image corpus. We then represent each image with a histogram of visual words. We extracted two codebooks, starting from two different random samples of points. We used soft assignment following Van Gemert et al. [34] using $\sigma = 90$. This descriptor was extracted using the executable publicly available from the University of Amsterdam [33] and from the VLFeat library [35]. We also extracted variations of the SIFT descriptor in different color spaces, namely RGB (red, green, blue), HSV (hue, saturation, value), and opponent channels.

The Fourier polar pyramid is similar to the curvelet feature, whereby each element of the feature vector represents the average of some region of Fourier-Mellin space. However, the regions are partitioned into a pyramid structure, introducing various degrees of scale and rotation invariance.

The third group of improvements applies to modeling in cases of data scarcity (opposite of large-scale conditions). Specifically, we implemented a variant of the IMARS system that can retrain unit models on the full 100% of training data, if no data subsampling was earlier employed for training of unit models. This helps boost the performance and ability of the unit models to generalize, while still supporting feature selection and ensemble model optimization. In cases of extreme data scarcity, ensembles of late fusion may be replaced with a single unit model trained over an early fusion of features.

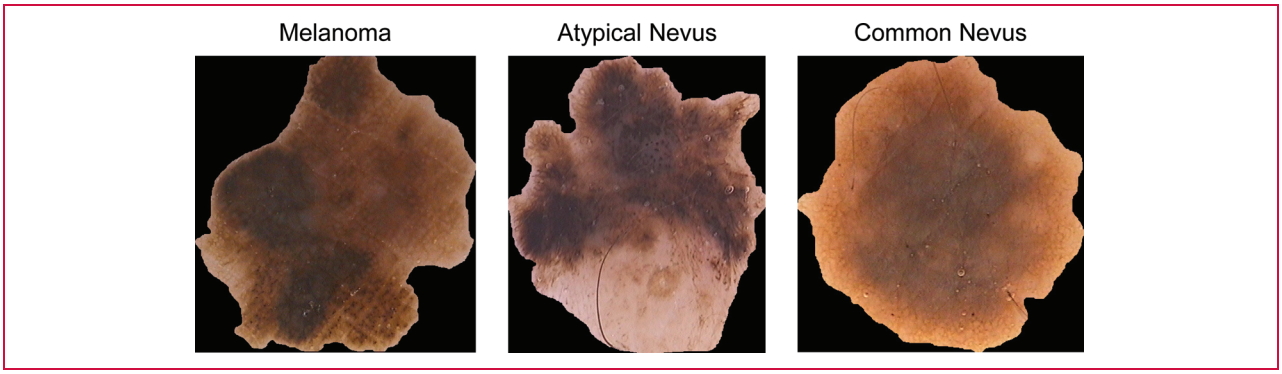

Figure 3

Examples from the PH2 dermoscopy dataset. "Melanoma" represents malignant disease lesions. Atypical nevi are lesions of suspicion, but non-malignant at time of imaging. Common nevi are benign lesions. Note the high degree of visual similarity between the classes.

In addition, the Synthetic Minority Oversampling Technique (SMOTE) has been implemented, which creates additional synthetic minority class training instances by taking linear combinations of data samples in feature space with their k-nearest neighbors.

The fourth and final improvement made to the system is the way model scores are mapped. Instead of the typical −infinity to +infinity of the SVM, we map scores to a logistic function that has been fit to the probabilities that a classifier score yields an instance of the positive class on a balanced dataset (computed during unit model cross-validation) [36].

Datasets

We demonstrate the performance of our visual analytics system in a collection of four medical imaging domains. All datasets were provided in de-identified form, intended for research purposes, according to HIPAA (Health Insurance Portability and Accountability Act) guidelines. In the first domain, we use a public medical image modality benchmark dataset. In the second domain, we utilize a collection of echocardiography video sequences of 340 patients, which involves 3 modes and 7 views (see "Echocardiography view and mode dataset"). In the third domain, we used two datasets of dermatology dermoscopy images that exhibit diseased states of melanoma and benign non-melanoma lesions. In the fourth domain, we aggregated multiple available datasets to create a collection of 158 categories covering various modalities, body regions, views, and disease states.

Standardized public ImageCLEF 2013 medical modality dataset

We utilized the ImageCLEF 2013 Modality Classification dataset [4], which contains 31 categories and is partitioned into fixed training and test datasets. The categories cover a wide variety of diagnostic medical images found in PubMed journal articles, across domains such as radiography (x-ray, computed tomography, magnetic resonance imaging, positron emission tomography, etc.), pathology

Table 2 Top 25 low-level features and resultant mean average precision (MAP) on ImageCLEF 2013 validation data partition.

MAP	Feature	Granularity
0.56492	SEMANTIC MODEL VECTOR	Mixed
0.5164	RGB-sift-VLFEAT-code2	Pyramid
0.50532	RGB-sift-VLFEAT-code1	Pyramid
0.50307	sift-VLFEAT-code2	Pyramid
0.50234	hsv-sift-VLFEAT-code1	Pyramid
0.49658	sift-VLFEAT-code1	Pyramid
0.481	hsv-sift-VLFEAT-code2	Pyramid
0.46803	LBP512-RGBH-59	Pyramid
0.44454	opponent-sift-AM-code2	Pyramid
0.44022	opponent-sift-AM-code1	Pyramid
0.43438	csiftAM-code1	Pyramid
0.4329	LBP512-RGBH-256	Pyramid
0.42609	csift-AM-code2	Pyramid
0.42309	siftAM-code2	Pyramid
0.42021	siftAM-code1	Pyramid
0.41924	LBP512-RGBH-256	Global
0.41422	lbp_histogram	Pyramid3
0.40978	mslbpgray59	Pyramid23
0.40541	rgbsiftAM_code2	Pyramid
0.39448	rgbsiftAM_code1	Pyramid
0.3889	hsvsiftAM_code2	Pyramid
0.38873	LBP320_gray	Pyramid
0.38148	mslbpcolorhue59	Global
0.37787	LBP320_gray	Grid7
0.3774	LBP320_gray	Grid7

Table 3 Experimental results for echocardiography view and mode detection.

Label	*Accuracy*	*Average precision*	*Sensitivity*	*Specificity*
Two-chamber	0.939	0.608	0.739	0.96
Three-chamber	0.911	0.696	0.768	0.94
Four-chamber	0.785	0.643	0.775	0.79
Five-chamber	0.849	0.017	0.168	0.90
CW-Doppler	0.999	0.955	1.00	1.00
PW-Doppler	0.997	0.847	1.00	1.00
Color Doppler	0.758	0.709	0.96	0.64
M-mode	1.00	1.00	1.00	1.00

(slide microscopy and electron microscopy), laboratory tests (chromatography gels), electrical signals (electrocardiograms), and visible light (endoscopy and dermoscopy). For the full accounting of categories, see **Table 1**. The training dataset contains 2,845 images, and the test set contains 2,582 images. Benchmark performance is measured by multiclass accuracy.

Echocardiography view and mode dataset

Our echo dataset consists of 340 patients and 2,158 echocardiographic sequences depicting a variety of cardiac diseases in patients including aneurysms (89), dilated cardiomyopathy (76), hypertrophies (78), and normal LV (left ventricle) size and function (448). Image frames have been extracted from video sequences, with the dataset involving a combination of mode and views: M-mode (MMOD), 2D or B-mode [which consists of four views of two-chamber (2CH), three-chamber (3CH), four-chamber (4CH), and five-chamber (5CH)], and Doppler echocardiography, which includes continuous wave (CW) Doppler (CWD), pulsed wave (PW) Doppler (PWD), and color flow Doppler (CFD). Therefore, we have eight categories/classes with total number of 83,381 images with very diverse distributions; M-mode (48), two chamber (7,524), three chamber (13,168), four chamber (27,954), five chamber (5,474), CW-Doppler (254), PW-Doppler (124), and color Doppler (28,835). **Figure 2** shows visual examples of these categories.

Dermoscopy disease datasets

For our experiments in dermoscopy melanoma recognition, we have utilized two datasets, summarized in the following subsections.

Pedro Hispano Hospital dataset

The first is the Pedro Hispano Hospital (PH2) dataset, containing a total of 200 images (40 instances of melanoma, and 160 instances of non-melanoma, including 79 atypical nevi). **Figure 3** shows example images from this dataset, emphasizing the high degree of similarity between some instances of the classes. Images are supplied with segmentations extracting the lesion from surrounding skin. The dataset is publicly available online, and has a reference standard measure of performance from prior literature.

ISIC dataset

The second is a dataset obtained through collaboration with the International Skin Imaging Collaboration (ISIC) [37]. This dataset includes 391 dermoscopy images of melanoma, and 2,314 dermoscopy images of benign lesions, a subset of which (225) are considered "near-miss" atypical lesions (visually similar to melanoma, as judged by medical professionals). The images of this dataset come without lesion segmentations. Therefore, for recognition of disease state, we simply analyzed regions defined by manually delineated bounding boxes around the areas of the skin lesions, in order to eliminate erroneous areas of the image that may influence recognition results.

Broad domain medical image dataset

The purpose of this dataset is to evaluate the performance of our ensemble algorithm when modeling a broad variety of medical images, covering modalities, body regions, views, and in some circumstances, disease states. In order to construct a dataset diverse enough to achieve this goal, we aggregated several publicly available datasets. In addition, we further augmented these data sources with annotated web search retrieval results to reduce the deficiencies in the existing data.

In total, we collected data for 158 medical imaging categories, containing 39,811 images. These categories were organized into a hierarchical taxonomy, ordered by modality, body region, view, and disease state. Datasets that were aggregated included the IRMA 2009 dataset [3], The Cancer Imaging Archives (TCIA) [38], the Japanese Society of Radiological Technology (JSRT) [39], and those acquired through collaboration with the ISIC. A full accounting of the categories and the number of positive exemplars in each category can be found in the data referenced in the Appendix.

For our experiments, the dataset was split into two partitions: 80% for model training and 20% as a held-out test set.

Table 4 Confusion matrix between echo concept classifiers, according to the Spearman Rank Correlation coefficient. Values above 0.25 are displayed in bold. Categories include M-mode (MMOD), two-dimensional or B-mode, which consists of four views of two-chamber (2CH), three-chamber (3CH), four-chamber (4CH), and five-chamber (5CH), and Doppler echocardiography, which includes continuous wave Doppler (CWD), pulsed wave doppler (PWD), and color flow Doppler (CFD).

	CFD	*MMOD*	*2CH*	*3CH*	*4CH*	*5CH*	*CWD*	*PWD*
CFD	**1.00**	0.04	**0.37**	0.19	**0.37**	0.11	**0.32**	0.13
MMOD	0.04	**1.00**	0.08	0.21	0.05	0.07	0.15	0.17
2CH	**0.37**	0.08	**1.00**	**0.35**	0.22	0.19	0.00	0.23
3CH	0.19	0.21	**0.35**	**1.00**	**0.59**	**0.36**	0.16	0.04
4CH	**0.37**	0.05	0.22	**0.59**	**1.00**	0.17	**0.40**	0.07
5CH	0.11	0.07	0.19	**0.36**	0.17	**1.00**	0.17	0.05
CWD	**0.32**	0.15	0.00	0.16	**0.40**	0.17	**1.00**	0.15
PWD	0.13	0.17	0.23	0.04	0.07	0.05	0.15	**1.00**

Table 5 Feature-level mean average precision (MAP) evaluated at full depth on the ISIC dataset Validation data partition, including clearly benign lesions.

MAP	*Feature*	*Granularity*
0.977	Lbp512-RGBHS-59	Pyramid23
0.9755	Lbp512-RGBHS-256	Pyramid
0.973	Lbp512-RGBHS-256	Pyramid23
0.965	Lbp512-RGBH-59	Pyramid23
0.962	Lbp512-RGBH-59	Pyramid
0.942	Lbp512-gray-59	Pyramid23
0.917	Lbp320-gray59	Pyramid23
0.813	Color-correlogram	Pyramid3
0.804	Color-wavelet	Pyramid3
0.756	Image-stats	Pyramid3
0.733	Image-type	Pyramid23
0.681	Gist	Pyramid
0.667	Edge-histogram	Pyramid23
0.661	Maxi-thumbnail-vector_global	Global
0.64566	Thumbnail-vector	Global
0.622	Mini-thumbnail-vector	Global
0.593	Color-histogram	Pyramid3
0.51	Wavelet-texture	Pyramid3
0.499	Fourier-polar	Pyramid

Table 6 Feature-level mean average precision (MAP) evaluated at full depth on ISIC dataset Validation data partition, excluding clearly benign lesions. Note: the absolute performance values are not comparable to those of Table 5, since the test set has changed in size and scope.

MAP	*Feature*	*Granularity*
0.968	Lbp512-RGBHS-256	Pyramid
0.964	Lbp512-RGBHS-256	Pyramid23
0.959	Lbp512-RGBHS-59	Pyramid23
0.9485	Lbp512-RGBH-59	Pyramid
0.944	Lbp512-RGBH-59	Pyramid23
0.942	Lbp512-gray-59	Pyramid23
0.938	Image-type	Pyramid23
0.935	Lbp320-gray59	Pyramid23
0.902	Color-wavelet	Pyramid3
0.899	Maxi-thumbnail-vector	Global
0.884	Thumbnail-vector	Global
0.881	Edge-histogram	Pyramid23
0.878	Image-stats	Pyramid3
0.878	Color-correlogram	Pyramid3
0.877	Mini-thumbnail-vector	Global
0.847	Gist	Pyramid
0.845	Curvelets	Pyramid
0.842	Color-histogram	Pyramid3
0.834	Wavelet-texture	Pyramid3

Experimental results

In the subsequent sections, we review the details and results for experiments performed on each of the four datasets.

ImageCLEF 2013 medical modality recognition

For the ImageCLEF 2013 medical modality task, we used our ensemble modeling system to train 1-vs-all classifiers for each of the categories. Multiclass decisions were made for each image by choosing the concept classifier with the maximum score.

Classifiers for ImageCLEF were trained using an algorithm variant for sparse data that implements a multi-stage retraining process: training data was first split into two sets of 50%. One set was used to train unit models, while the

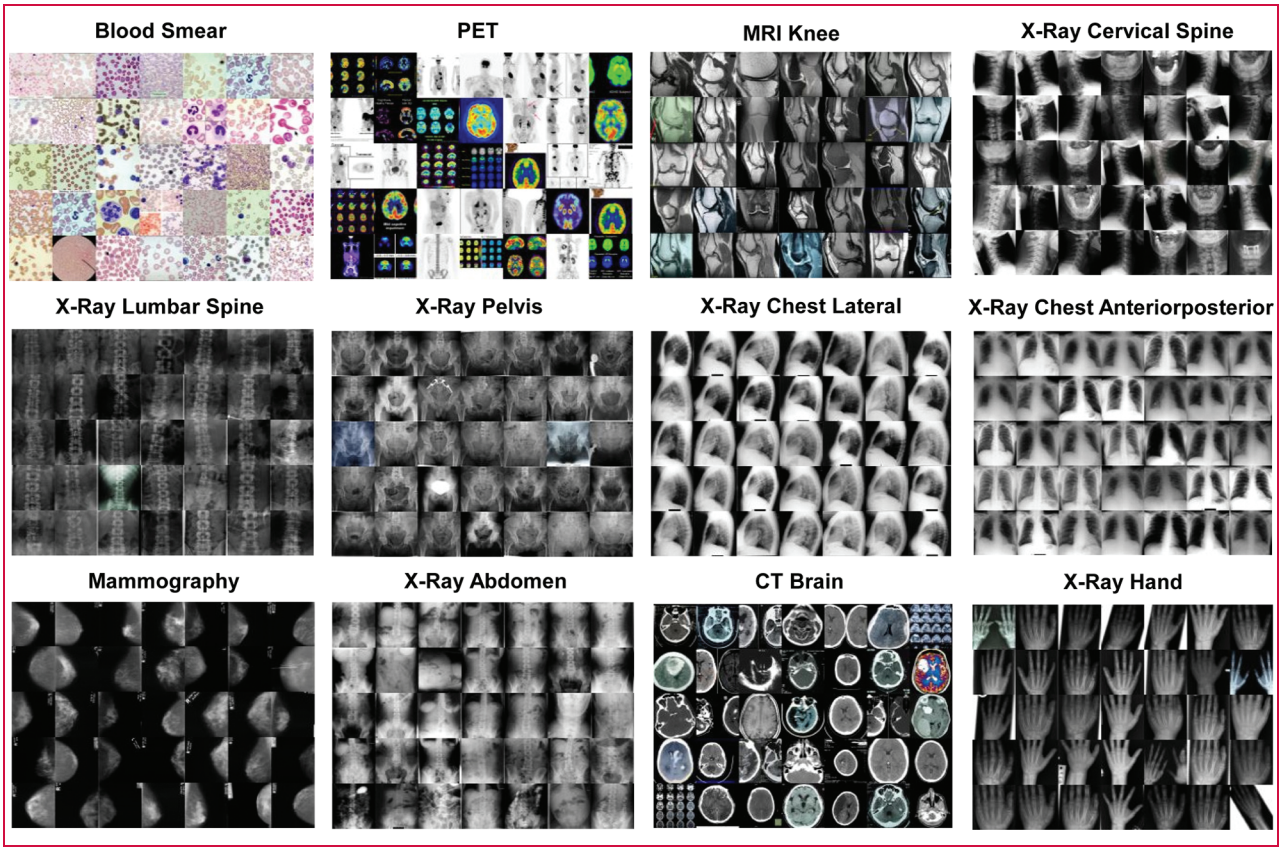

Figure 4

Example visual retrieval results from the Broad Domain Medical image dataset.

other set was used to compute ensembles of unit models for each category. Subsequently, after ensemble learning, unit models selected in each ensemble were retrained on 100% of the available data. This allowed us to maximize the utility of available data, as some categories in the ImageCLEF 2013 dataset have very few numbers of exemplars.

Additionally, we performed experiments utilizing a semantic model vector, a descriptor in which each of the 120 dimensions corresponds to the score of a model trained on a broad domain category (none of the semantic model vector categories overlaps with ImageCLEF ones). We compare system performance with and without this additional feature.

In summary, our ensemble modeling approach yields 81.17% multiclass accuracy without use of a 120-dimensional broad domain medial semantic model vector, and 82.2% with the use of this high-level semantic model vector. Both these performance levels set a new state-of-art.

Individual performance of the top 25 features is shown in **Table 2**. The semantic model vector was the best performing single feature, with SIFT variants and LBP following.

Echocardiography view and mode recognition

In this task, the problem presents with highly unbalanced data. Some of the categories have a very small number of samples, others have tens of thousands images. The data was split into 80% for training and 20% for held-out test. We split the images in a way that ensures all images of every patient either belongs to test set or training set. Then we use one-vs-all classifiers approach to train our ensemble model per each category. The final predicting label is decided by considering the maximum score.

Resultant multiclass accuracy was 90.48%. Detailed performance metrics per category are shown in **Table 3**. The confusion matrix between the classifiers according to their Spearman Rank Correlation on the test set is shown in **Table 4**. The most correlated classifiers were 3-chamber and 4-chamber views. Inspection of Figure 2 confirms that visually these categories are among the most similar.

Dermoscopy disease recognition

In the following subsections, we review experiments performed on the two dermatology datasets described in the section "Dermoscopy disease datasets."

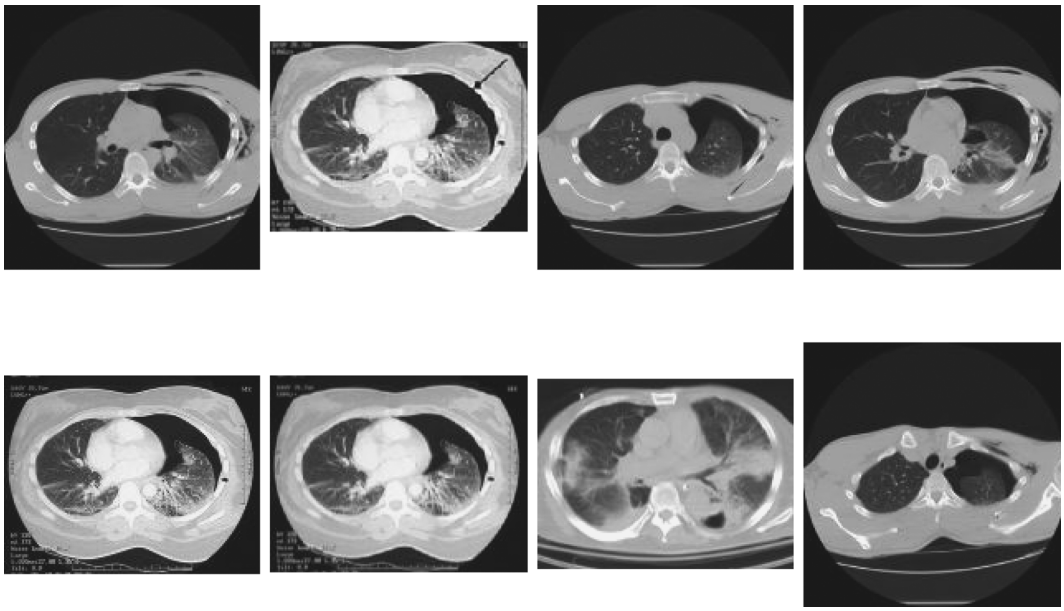

Figure 5

Top-scoring example retrieval results for disease state "Pneumothorax" (lung collapse) for the CT imaging modality. One can clearly identify the collapsed left lung in the images (chest cavity on right side of image with black air gap as a result of the lung collapse).

PH2 dataset

Because of the extreme scarcity of exemplars in the PH2 dataset, ensemble strategies are not effective, as not enough data is present to perform both unit model training and ensemble fusion. Therefore, we used a variant of IMARS that trains a single unit model with early fusions of features. 2-fold cross validation is still used for logistic score fitting in order to address data imbalance. Evaluations are carried out in accordance with prior literature, utilizing a leave-one-out strategy: one example is left out of training, while models are trained on the remaining data and used to make a judgment of the sample left out. This is repeated until all samples have been left out.

Our experimental approach was to start with a simple feature to describe color, and iteratively add features that better describe texture, or interactions between color and texture. We expected to see an improvement in performance as additional image statistics are involved in the training process, until the feature combinations become of sufficiently high dimension that overfitting starts to occur. Indeed, this is the pattern that our experiments show; however, before saturation occurs, state-of-art performance was obtained.

In total, we performed four experiments. In the first, we utilized the color histogram feature at global granularity. Resultant sensitivity and specificity is 0.675 and 0.9062, respectively, with an average precision (AP) at full depth of 0.743. In the second, we concatenated color and edge histograms at global granularities. Performance improved to 0.8 and 0.9375 sensitivity and specificity, respectively, with an average precision at full depth of 0.88. In the third, we concatenated color, edge, and color LBP histograms (59 bins), all at global granularities. Performance capped at 0.9 and 0.9 sensitivity and specificity, respectively, with an average precision of 0.927. Using a threshold where sensitivity is fixed to a value of 0.93 as reported in prior literature, this result improves state-of-art by 4% in specificity (0.88 specificity versus 0.84 in prior reports [22]).

In the fourth and final experiment, we continued to concatenate additional features, including both "image type" and "image stats" feature vectors [3]. These features measure global image statistics, such as mean saturation, hue entropy, variance and switches, quantized color entropy and switches, variance, minimum value, maximum value, mean, median, standard deviation, central moments, average energy of the first level of 2D wavelet decomposition subbands, skin color, and number of unique colors in quantized color space. However, we found performance to decrease, likely due to overfitting of the small dataset from the feature vector becoming too large. Sensitivity and specificity reduced to 0.9 and 0.888, respectively, and AP fell to 0.922.

ISIC dataset

As the ISIC dataset is an order of magnitude larger than the PH2 data, we changed back to the ensemble modeling

Table 7 Full accounting of Broad Domain Medical 158 concept categories, the number of positive exemplars per category, and the AP evaluated at full recall on the 20% held-out test set.

Concept	*No. of exemplars*	*AP*
CHART	790	0.966
VIS_DERM_CPWI	2,356	0.993
VIS_DERM_CPWI_ICLEF2013-DVDM-UNKNOWN	80	0.408
VIS_DERM_CPWI_MELANOMA	381	0.844
VIS_DERM_CPWI_NON-MELANOMA	1,895	0.993
VIS_DERM_CPWI_NON-MELANOMA_CHILDREN	1,669	0.997
VIS_DERM_CPWI_NON-MELANOMA_OTHER	226	0.747
CT	12,566	1
CT_BRAIN	169	0.894
CT_CHEST	12,397	1
CT_CHEST_LUNGCANCER	120	0.437
CT_CHEST_NORMAL	12,081	1
CT_CHEST_PNEUMOTHORAX	196	0.843
DERMATOLOGY	6,067	0.999
DERMOSCOPY_CROPPED	1,006	0.993
DERMOSCOPY_WHOLEIMAGE	2,705	0.997
SM_DM	227	0.705
SM_DM_EL	52	0.441
SM_DM_FL	34	0.594
SM_DM_LI	93	0.779
SM_DM_TR	48	0.558
VIS_DERM_DSCPY-CROP_MELANOMA	391	0.839
VIS_DERM_DSCPY-CROP_NON-MELANOMA_CHILDREN	390	0.933
VIS_DERM_DSCPY-CROP_NON-MELANOMA_OTHER	225	0.711
VIS_DERM_DSCPY-CROP_NON-MELANOMA	615	0.926
VIS_DERM_DSCPY-WI_MELANOMA	391	0.329
VIS_DERM_DSCPY-WI_NON-MELANOMA_CHILDREN	2,155	0.998
VIS_DERM_DSCPY-WI_NON-MELANOMA_OTHER	159	0.243
VIS_DERM_DSCPY-WI_NON-MELANOMA	2,314	0.988
DX	16,848	0.999
DX_APPENDAGE	5,811	0.986
DX_APPENDAGE_ARM	3,005	0.924
DX_APPENDAGE_ARM_ELBOW	474	0.872
DX_APPENDAGE_ARM_ELBOW_AP	199	0.804
DX_APPENDAGE_ARM_ELBOW_LAT	274	0.843
DX_APPENDAGE_ARM_FOREARM	198	0.699
DX_APPENDAGE_ARM_FOREARM_AP	91	0.534
DX_APPENDAGE_ARM_FOREARM_LAT	107	0.608
DX_APPENDAGE_ARM_HAND	1,452	0.917
DX_APPENDAGE_ARM_HAND_FINGER	189	0.49
DX_APPENDAGE_ARM_HAND_WHOLE	989	0.886
DX_APPENDAGE_ARM_HAND_WHOLE_AP	860	0.86

Table 7 *(Continued.)* Full accounting of Broad Domain Medical 158 concept categories, the number of positive exemplars per category, and the AP evaluated at full recall on the 20% held-out test set.

DX_APPENDAGE_ARM_HAND_WHOLE_OBL	129	0.879
DX_APPENDAGE_ARM_HAND_WRIST	273	0.871
DX_APPENDAGE_ARM_HAND_WRIST_AP	135	0.895
DX_APPENDAGE_ARM_HAND_WRIST_LAT	138	0.786
DX_APPENDAGE_ARM_SHOULDER	792	0.94
DX_APPENDAGE_ARM_UPPER	88	0.295
DX_APPENDAGE_LEG	2,806	0.925
DX_APPENDAGE_LEG_ANKLE	450	0.967
DX_APPENDAGE_LEG_ANKLE_AP	242	0.973
DX_APPENDAGE_LEG_ANKLE_LAT	208	0.925
DX_APPENDAGE_LEG_FOOT	827	0.871
DX_APPENDAGE_LEG_FOOT_AP	494	0.813
DX_APPENDAGE_LEG_FOOT_LATERAL	151	0.573
DX_APPENDAGE_LEG_FOOT_OBLIQUE	181	0.663
DX_APPENDAGE_LEG_KNEE	1,042	0.927
DX_APPENDAGE_LEG_KNEE_PATELLA	123	0.958
DX_APPENDAGE_LEG_KNEE_WHOLE	918	0.911
DX_APPENDAGE_LEG_KNEE_WHOLE_AP	550	0.908
DX_APPENDAGE_LEG_KNEE_WHOLE_LAT	368	0.871
DX_APPENDAGE_LEG_LOWER	245	0.563
DX_APPENDAGE_LEG_LOWER_AP	156	0.501
DX_APPENDAGE_LEG_LOWER_LAT	88	0.322
DX_APPENDAGE_LEG_UPPER	241	0.527
DX_APPENDAGE_LEG_UPPER_AP	168	0.569
DX_APPENDAGE_LEG_UPPER_LAT	72	0.407
DX_CRANIUM	1,338	0.97
DX_CRANIUM_NOSE	406	0.904
DX_CRANIUM_NOSE_LAT	64	0.739
DX_CRANIUM_NOSE_OCCIPITOFRONTAL	342	0.926
DX_CRANIUM_WHOLE	931	0.978
DX_CRANIUM_WHOLE_AP	436	0.963
DX_CRANIUM_WHOLE_FRONTOCCIPITAL	50	0.759
DX_CRANIUM_WHOLE_LAT	445	0.946
DX_HIP_JOINT_AP_ARTIFICIAL	112	0.616
DX_HIP_JOINT_AP_NATURAL	167	0.484
DX_HIP_JOINT_LAT_ARTIFICIAL	26	0.165
DX_HIP_JOINT_LAT_NATURAL	68	0.67
DX_TORSO	9,698	0.995
DX_TORSO_ABDOMEN	467	0.833
DX_TORSO_ABDOMEN_AP	247	0.882
DX_TORSO_ABDOMEN_BARIUMSWALLOW	177	0.711
DX_TORSO_ABDOMEN_UPPER	43	0.863
DX_TORSO_BREAST	335	0.984
DX_TORSO_BREAST_LEFT	171	0.971
DX_TORSO_BREAST_LEFT_AXIAL	85	0.902
DX_TORSO_BREAST_LEFT_OBL	86	0.921
DX_TORSO_BREAST_RIGHT	164	0.97
DX_TORSO_BREAST_RIGHT_AXIAL	80	0.902

Table 7 *(Continued.)* Full accounting of Broad Domain Medical 158 concept categories, the number of positive exemplars per category, and the AP evaluated at full recall on the 20% held-out test set.

DX_TORSO_BREAST_RIGHT_OBL	84	0.875
DX_TORSO_CHEST	6,308	0.998
DX_TORSO_CHEST_FULL	6,308	0.999
DX_TORSO_CHEST_FULL_AP	5,043	0.999
DX_TORSO_CHEST_FULL_AP_LUNGCANCER	594	0.931
DX_TORSO_CHEST_FULL_AP_NORMAL	383	0.291
DX_TORSO_CHEST_FULL_AP_PNEUMONIA	262	0.261
DX_TORSO_CHEST_FULL_AP_PNEUMOTHORAX	217	0.364
DX_TORSO_CHEST_FULL_AP_UNKNOWN	3,587	0.998
DX_TORSO_CHEST_FULL_LAT	1,265	0.996
DX_TORSO_CHEST_FULL_LAT_LUNGCANCER	202	0.965
DX_TORSO_CHEST_FULL_LAT_NORMAL	21	0.109
DX_TORSO_CHEST_FULL_LAT_UNKNOWN	1,042	0.992
DX_TORSO_HIP	821	0.945
DX_TORSO_HIP_JOINT	373	0.863
DX_TORSO_HIP_JOINT_AP	279	0.712
DX_TORSO_HIP_JOINT_LAT	94	0.806
DX_TORSO_HIP_PELVIS	447	0.98
DX_TORSO_HIP_PELVIS_ARTIFICIAL	94	0.779
DX_TORSO_HIP_PELVIS_NATURAL	353	0.944
DX_TORSO_SPINE	1,767	0.978
DX_TORSO_SPINE_CERVICAL	798	0.963
DX_TORSO_SPINE_CERVICAL_AP	294	0.951
DX_TORSO_SPINE_CERVICAL_LAT	503	0.924
DX_TORSO_SPINE_LUMBAR	668	0.945
DX_TORSO_SPINE_LUMBAR_AP	365	0.901
DX_TORSO_SPINE_LUMBAR_LAT	303	0.911
DX_TORSO_SPINE_THORACIC	300	0.979
DX_TORSO_SPINE_THORACIC_AP	144	0.966
DX_TORSO_SPINE_THORACIC_LAT	156	0.959
ECG	1,131	0.994
ECG_FIBRILLATION	734	0.714
ECG_NORMAL	397	0.602
VIS_ICLEF2013-DVEN	65	0.744
VIS_ICLEF2013-DVOR	71	0.595
CHART_ICLEF2013-GCHE	65	0.862
CHART_ICLEF2013-GFIG	107	0.872
CHART_ICLEF2013-GFLO	100	0.593
CHART_ICLEF2013-GGEL	70	0.728
CHART_ICLEF2013-GGEN	90	0.638
CHART_ICLEF2013-GHDR	47	0.691
CHART_ICLEF2013-GMAT	21	0.206
VIS_ICLEF2013-GNCP	97	0.748
CHART_ICLEF2013-GPLI	29	0.833
CHART_ICLEF2013-GSCR	95	0.576
CHART_ICLEF2013-GSYS	96	0.27
CHART_ICLEF2013-GTAB	70	0.593
MR	1,276	0.929

Table 7 *(Continued.)* Full accounting of Broad Domain Medical 158 concept categories, the number of positive exemplars per category, and the AP evaluated at full recall on the 20% held-out test set.

MR_APPENDAGE	416	0.855
MR_BRAIN	518	0.877
MR_BRAIN_AXIAL	333	0.739
MR_BRAIN_AXIAL_COLLAGE	134	0.605
MR_BRAIN_AXIAL_SINGLE	198	0.785
MR_BRAIN_SAGITTAL	185	0.876
MR_HIP	125	0.563
MR_KNEE	416	0.847
MR_SPINE	217	0.863
PET	304	0.897
PET_BW	166	0.823
PET_COLOR	138	0.896
SM	645	0.949
SM_BLOODSMEAR	418	0.967
SM_BLOODSMEAR_NORMAL	360	0.78
SM_BLOODSMEAR_SICKLECELL	58	0.37
US	490	0.965
US_CARDIAC	233	0.777
US_FETUS	257	0.82
VIS	6,300	0.999

algorithm. For our experiments, we studied two variants of the ISIC dataset. The first includes data from lesions that are clearly benign, and the second excludes data from clearly benign lesions, which may result in a more difficult task. For both experiments, 80% of data was used for training, and 20% was used to test the algorithm.

Experiments with clearly benign lesions involved 391 images of melanoma, 225 images of atypical lesions, and 2,536 clearly benign lesions. Resultant AP at full depth was 0.967. At the cutoff threshold of 0.5 (logistically normalized SVM scores), we measured a sensitivity of 0.987 and specificity of 0.9482. At threshold with fixed sensitivity values of 0.99, the experiment yields a specificity value of 0.9445.

Excluding clearly benign lesions (391 melanoma, 225 atypical), resultant AP at full depth was 0.983. At the cutoff threshold, trained models produced a sensitivity of 0.846 and a specificity of 0.9375. At a threshold with fixed sensitivity values of 0.99, the experiment yielded a specificity value 0.594.

Table 5 shows the average precision of our low-level features on the Validation partition of the ISIC dataset including clearly benign lesions (feature types and granularities are as specified in the section "Visual modeling approach"). **Table 6** presents the same analysis for the dataset excluding benign lesions. LBP features have specified the standard square size that images are rescaled to 320 × 320 or 512 × 512, the color channels (*gray*, *R*ed, *G*reen, *B*lue, *H*ue, *S*aturation), and the number of bins in the histogram (59 or 256). Granularities are appended to the end (see the section "Visual modeling approach"). For review, "global" is simply feature extracted from the whole image (1 × 1). "Pyramid" is whole image (1 × 1) concatenated with quarters (2 × 2). "Pyramid23" is whole image (1 × 1) concatenated with quarters (2 × 2), concatenated with ninths (3 × 3).

Clearly, color LBP features dominate. Using 256 bins as opposed to 59 bins appears to yield no benefit. Additional pyramid levels add marginal benefit. Increasing resolution of image rescaling before feature extraction brings improved performance.

To demonstrate the importance of the ensemble fusion involving multiple features, we re-ran the experiment excluding clearly benign lesions with the single top performing feature in that scenario. Resultant threshold sensitivity was 0.872, with a specificity of 0.901. At a threshold with sensitivity of 0.99, specificity is reduced to 0.438.

Broad domain medical image recognition

In our broad domain medical image recognition experiments, we trained 158 one-vs-all classifiers for each of the categories in the dataset. The size of the data subsamples, or "bags," for each feature were up to 5,000 positive exemplars and 5,000 negative examples, and 10 bags per feature to cover the entire data space. Evaluation is carried out on the held-out 20% dataset.

Given that concepts are defined in a hierarchical fashion, positive and negative exemplars may be sampled from one or more concepts in the dataset (children or siblings/cousins). We employed two strategies of sampling data across multiple concepts in these scenarios: 1) concept proportions in the data subsample equals proportions seen in the training data and 2) concept proportions in the data subsample is equalized so that each category receives as equal a representation as possible, while still filling the bag.

Example visual retrieval results are depicted in **Figure 4**. More detailed results for an example disease state, pneumothorax, within the "CT" modality and "Chest" body region are shown in **Figure 5**. In summary, the ensemble model strategy using the second equal representation sampling yielded the highest mean average precision (MAP) of 0.792, slightly above the first sampling strategy, which yielded an MAP of 0.785. Average accuracy was 0.984, with an average sensitivity of 0.947 and specificity of 0.984. Correlating concept classifier performance, in terms of AP, with the tree-depth of the concept in the hierarchy yielded no significant correlation ($R^2 = 0.024$), suggesting the general to specific ordering of concepts was not predictive of resultant classifier quality. Individual classifier performance metrics for each category can be found in the Appendix.

In total, this task required training over 26,000 SVMs (over 78,000 counting cross-fold validation) and learning 158 ensembles. Utilizing our hardware resources (see the section "Visual modeling approach"), training took 9.5 hours. Once features are pre-extracted, scoring the models on the test set (8,000 images) took 12 minutes on 700 cores. This translates to 0.4 seconds per instance, per classifier, per core. Feature extraction required just under an additional 2.5 seconds per core, yielding a total evaluation time of approximately 3 seconds per image, per classifier, per core.

Clearly, the ability to arbitrarily scale to 700 cores made the large-scale experiment feasible. On a single core, the same experiment would have taken over 290 days to train and 6 days to evaluate.

Scaling our experiment even further, we studied how performance might be improved if we optimized SVM kernel selection for each unit model trained. For this experiment, four kernels were trained for each unit model, and the best performing selected for use. The four kernels spanned histogram intersection and chi-squared kernels, with varying values of misclassification costs. Resultant MAP increased to 0.801, from 0.792, an improvement of approximately 1%. Since optimizing over four kernels required training four times as many SMVs (over 100,000), the cost/benefit ratio is clearly very high in comparison to using a single SVM kernel.

Conclusion

We presented a flexible and scalable modeling system for recognition of medical image categories. Performance was evaluated in the following four contexts: 1) modality, 2) echocardiography view and mode, 3) melanoma, and 4) broad medical domain categories. In the first context, our system achieved state-of-art performance of 82.2% on the public ImageCLEF 2013 benchmark dataset. In the second, the system achieved performance of 90.48% multiclass accuracy. In the third, we achieved state-of-art performance on a small public benchmark dataset of 200 images and demonstrated an ability to generalize to a larger dataset obtained through collaboration with the ISIC. In the fourth and last context, we studied the system's ability to scale to

Table 8 List of acronyms for the Broad Domain Medical dataset.

Abbreviation	*Meaning*
AP	Anterior-posterior view
BW	Grayscale color spectrum
CPWI	Clinical Photography (Whole Image)
CT	Computed Tomography
DERM	Dermatology
DM	Dermoscopy (ImageCLEF)
DSCPY-CROP	Dermoscopy (Cropped ISIC)
DSCPY-WI	Dermoscopy (Whole Image ISIC)
DVDM-UNKNOWN	Dermatology (Undiagnosed)
DVEN	Endoscopy
DVOR	Other Organs
DX	Digital X-Ray
ECG	Electrocardiogram
EL	Electron Microscopy
FL	Fluorescent Microscopy
GCHE	General (chemical structure)
GFIG	General (figure)
GFLO	General (flow diagram)
GGEL	General (chromatography gel)
GGEN	General (gene sequence)
GHDR	General (hand drawn)
GMAT	General (math equation)
GNCP	General (non-clinical photograph)
GPLI	General (program listing)
GSCR	General (screenshot)
GSYS	General (system diagram)
GTAB	General (table)
ICLEF2013	ImageCLEF2013 Inherited Concept
LAT	Lateral view
LI	Light microscopy
MR	Magnetic Resonance
OBL	Oblique
PET	Positron emission tomography
SM	Slide microscopy
TR	Transmission microscopy
US	Ultrasound
VIS	Visible light spectrum images

158 medical concepts covering broad and specific categories of modality, body region, view point, and disease. Resultant performance yielded average sensitivity and specificity of 0.95 and 0.98, respectively.

In summary, the proposed ensemble visual modeling system has been shown to be an effective tool for a broad range of medical image categories. Further research is warranted to study combining the framework with libraries of more specialized techniques and algorithms targeted for finely detailed local analysis and quantitative measurements in specific modalities, body regions, and views, which have the potential to improve the ability to extract relevant evidence to support the diagnosis of more disease states.

Appendix I: Broad domain 158 categories

The categories of our curated dataset, and the number of instances contained within each, are displayed in **Table 7**. The full path in the hierarchy is designated with underscores, which indicate a new branch. Abbreviations are outlined in **Table 8**.

**Trademark, service mark, or registered trademark of National Library of Medicine, MELA Sciences, Apache Software Foundation in the United States, other countries, or both.

References

1. Frost & Sullivan. (2004). 2004 healthcare storage report, Mountain View, CA, USA. [Online]. Available: https://www.emc.com/collateral/analyst-reports/4_fs_wp_medical_image_sharing_021012_mc_print.pdf
2. National Center for Health Statistics. [Online]. Available: http://www.cdc.gov/nchs/data/databriefs/db105.pdf
3. T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein, "The IRMA code for unique classification of medical images," in *Proc. SPIE*, 2003, vol. 5033, pp. 109–117. [Online]. Available: http://ganymed.imib.rwth-aachen.de/irma/index_en.php
4. B. Caputo, H. Muller, B. Thomee, M. Villegas, R. Paredas, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. M. Gomez, I. G. Varea, and M. Cazorla, "ImageCLEF 2013: The Vision, the Data and the Open Challenges," in *Proc. Inf. Access Eval., Multilinguality, Multimodality, Vis.*, vol. 8138, *Lecture Notes in Computer Science*, 2013, pp. 250–268. [Online]. Available: http://imageclef.org/2013/medical
5. U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X-ray Categorization and Retrieval on the Organ and Pathology Level, Using Patch-Based Visual Words," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 733–746, Mar. 2011.
6. S. K. Antani, T. M. Deserno, L. R. Long, M. O. Guld, L. Neve, and G. R. Thoma, "Interfacing Global and Local CBIR Systems for Medical Image Retrieval," in *Proc Workshop Med. Imag. Res. (Bildverarbeitung fur die Medizin)*, Mar. 2007, pp. 166–171.
7. S. K. Antani, T. M. Deserno, L. R. Long, and G. R. Thoma, "Geographically Distributed Complementary Content-Based Image Retrieval Systems for Biomedical Image Informatics," *Studies Health Technol. Informat.*, vol. 129, pp. 493–497, 2007.
8. L. Cao, N. Codella, J. Connell, M. Merler, Q. Nguyen, S. Pankanti, and J. Smith, "IBM Multimedia Analytics @ ImageCLEF2013," in *Proc. ImageCLEFmed Med. Image Retrieval Workshop*, 2013. [Online]. Available: http://www.imageclef.org/system/files/IBMimageCLEF13.pdf
9. D. Beymer and T. Syeda-Mahmood, "Cardiac disease recognition in echocardiograms using spatio-temporal statistical models," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 4784–4788.
10. Y. Qian, L. Wang, C. Wang, and X. Gao, "The synergy of 3D SIFT and sparse codes for classification of viewpoints from echocardiogram videos," in *Proc. Med. Content-Based Retrieval Clinical Decision Support*, vol. 7723, *Lecture Notes in Computer Science*, 2013, pp. 68–79.
11. S. Ebadollahi, S. Chang, and H. Wu, "Automatic view recognition in echocardiogram videos using parts-based representation," in *Proc. CVPR*, 2004, pp. 2–9.
12. S. K. Zhou, J. H. Park, B. Georgescu, D. Comaniciu, C. Simopoulos, and J. Otsuki, "Image-based multiclass boosting and echocardiographic view classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1559–1565.
13. M. Otey, J. Bi, S. Krishna, B. Rao, J. Stoeckel, A. S. Katz, J. Han, and S. Parthasarathy, "Automatic view recognition for cardiac ultrasound images," in *Proc. MICCAI: Int. Workshop Comput. Vis. Intravascular Intracardiac Imag.*, 2006, pp. 187–194.
14. J. Park, S. Zhou, C. Simopoulos, J. Otsuki, and D. Comaniciu, "Automatic cardiac view classification of echocardiogram," in *Proc. ICCV*, 2007, pp. 1–8.
15. D. Beymer, T. Syeda-Mahmood, and F. Wang, "Exploiting spatio-temporal information for view recognition in cardiac echo videos," in *Proc. IEEE Comput. Soc. Workshop MMBIA*, 2008, pp. 1–8.
16. R. Kumar, F. Wang, D. Beymer, and T. Syeda-Mahmood, "Echocardiogram view classification using edge filtered scale-invariant motion features," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2009, pp. 723–730.
17. B. González, F. Valdez, P. Melin, and G. Prado-Arechiga, "Echocardiogram image recognition using neural networks," *Studies Comput. Intell.*, vol. 547, pp. 427–436, 2014.
18. H. Wu, D. M. Bowers, T. T. Huynhy, and R. Souvenir, "Echocardiogram view classification using low-level features," in *Proc. IEEE 10th Int. Symp. Biomed. Imag., Nano Macro*, 2013, pp. 752–755.
19. D. Agarwal, K. S. Shriram, and N. Subramanian, "Automatic view classification of echocardiograms using Histogram of Oriented Gradients," in *Proc. IEEE 10th ISBI*, Apr. 7–11, 2013, pp. 1368–1371.
20. SkinCancer.org. [Online]. Available: http://www.skincancer.org/
21. A. R. A. Ali and T. M. Deserno, "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data," in *Proc. SPIE*, Feb. 2012, vol. 8318, Art. ID. 831811.
22. C. Barata, M. Ruela, M. Francisco, T. Mendonca, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014.
23. R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border and wavelet-based texture analysis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1239–1252, Nov. 2012.
24. R. Garnavi, M. Aldeen, M. E. Celebi, S. Finch, and G. Varigos, "Border detection in dermoscopy images using hybrid thresholding on optimized color channels," *Comput. Med. Imag. Graphics*, vol. 35, no. 2, pp. 105–115, Mar. 2011.
25. Melafind Product Insert. [Online]. Available: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/GeneralandPlasticSurgeryDevicesPanel/UCM233837.pdf
26. Y. Liu, A. An, and X. Huang, "Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles," in *Proc. Adv. Knowl. Discovery Data Mining*, vol. 3918, *Lecture Notes in Computer Science*, 2006, pp. 107–118.
27. P. Kang and S. Cho, "EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems," in *Proc. Neural Inf. Process.*, vol. 4232, *Lecture Notes in Computer Science*, 2006, pp. 837–846.
28. R. Yan, M. O. Fleury, M. Merler, A. Natsev, and J. R. Smith, "Large-scale multimedia semantic concept modeling using robust

subspace bagging and MapReduce," in *Proc. 1st ACM Workshop Large-Scale Multimedia Retrieval Mining*, 2009, pp. 35–42.
29. N. Codella, A. Natsev, G. Hua, M. Hill, L. Cao, L. Gong, and J. R. Smith, "Video event detection using temporal pyramids of visual semantics with Kernel optimization and model subspace boosting," in *Proc. IEEE ICME*, 2012, pp. 747–752.
30. T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with LBP," in *Proc. Lecture Notes Comput. Sci.*, 2004, vol. 3021, pp. 469–481, Springer.
31. C. Zhu, C. Bichot, and L. Chen, "Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition," in *Proc. ICPR*, 2010, pp. 3065–3068.
32. E. Candès, L. Demanet, D. Donoho, and L. Ying. (2006). Fast Discrete Curvelet Transforms. *Multiscale Model. Simul.* [Online]. *5(3)*, pp. 861–899. Available: http://www.curvelet.org/
33. L. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
34. J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual Word Ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
35. A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1469–1472. [Online]. Available: http://www.vlfeat.org/
36. J. R. Kender, "Separability and refinement of hierarchical semantic video labels and their ground truth," in *Proc. IEEE ICME*, 2008, pp. 673–676.
37. International Skin Imaging Collaboration Website. [Online]. Available: http://www.isdis.net/index.php/projects/40-dec-2013-update
38. The Cancer Imaging Archive (TCIA). [Online]. Available: http://cancerimagingarchive.net/
39. Japanese Society of Radiological Technology. [Online]. Available: http://www.jsrt.or.jp/jsrt-db/eng.php

Received May 1, 2014; accepted for publication May 30, 2014

Mani Abedini *IBM Research - Australia, Carlton, Victoria 3053 Australia (mabedini@au1.ibm.com).* Dr. Abedini is a post-doctorate researcher in the Health Care Analytic team at the IBM Australia - Research Laboratories. He received a Ph.D. degree from the Department of Computing and Information Systems (CIS) at the University of Melbourne in 2014. His major research interests are big data analysis, machine learning, soft computing, and software development [including parallel and distributed computing on high-performance computing and GPGPU (general-purpose computing on graphics processing units)]. He joined IBM in 2013, initially as a research intern and later as a full-time researcher. During this time, he actively collaborates in healthcare analytic and life science projects.

Noel C. F. Codella *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (nccodell@us.ibm.com).* Dr. Codella is a Research Staff Member in the Multimedia Analytics Group in the Cognitive Computing organization at the IBM T. J. Watson Research Center. He received his B.S. degree in computer science from Columbia University in 2004, his M.Eng. degree in computer science from Cornell University in 2005, and his Ph.D. degree in physiology, biophysics, and systems biology from the Weill Cornell Medical College in 2010. Dr. Codella's thesis work included cardiac magnetic resonance imaging (MRI) free-breathing data acquisition, parallel MRI image reconstruction, and cardiac segmentation with functional analysis. He joined the IBM T. J. Watson Research Center in December 2010 as a Postdoctoral Researcher and later as Research Staff Member in October 2011. His research at IBM is focused on large-scale machine learning for visual analytics. He is the recipient of two IBM Invention Achievement Awards and an IBM Eminence and Excellence Award, and he is co-recipient of an IBM Research Division Award. He has over 25 publications, 13 conference abstracts, and nine patent filings. Dr. Codella is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM).

Jonathan H. Connell *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (jconnell@us.ibm.com).* Dr. Connell received his Ph.D. degree in artificial intelligence from MIT and then went to work at the IBM T. J. Watson Research Center. His projects include robot navigation, reinforcement learning, natural language processing, audio-visual speech recognition, video browsing, fingerprint identification, iris recognition, and cancelable biometrics. He has done extensive work in real-world computer vision including recognizing items in retail stores, object detection for video surveillance, and vehicle spotting for automotive controls. Most recently, he has developed a multi-modal instructional dialog system for use with speech-guided eldercare mobile robots. In addition, he has taught in the Psychology Department at Vassar College, is an Institute of Electrical and Electronics Engineers (IEEE) Fellow and the author of three books, and holds 48 U.S. patents.

Rahil Garnavi *IBM Research - Australia, Carlton, Victoria 3053 Australia (rahilgar@au1.ibm.com).* Dr. Garnavi received her B.Sc. degree in software engineering in 2003 and her M.Sc. degree in artificial intelligence in 2005. She completed her Ph.D. in 2011 at the University of Melbourne, Department of Electrical and Electronic Engineering, developing a computer-aided diagnostic system for melanoma. She joined IBM Research - Australia in June 2011 as a Research Scientist. She has been leading the multimedia analytics team since then, actively driving computer vision related projects in the healthcare domain and beyond. She has also held an honorary research fellowship position in the Department of Computing and Information Systems at The University of Melbourne, and she has mentored students. Dr. Garnavi has authored numerous research papers in image analytics presented in conferences and published in peer-reviewed journals, and also holds four patents. Some of her work experience in industry prior to her Ph.D. includes business analysis and data modeling, as well as software design and development. Her research interests involve image and video analytics, machine learning, and cognitive computing. Dr. Garnavi is a member of Institute of Electrical and Electronics Engineers (IEEE).

Michele Merler *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (mmerler@us.ibm.com).* Dr. Merler is a Research Staff Member in the Multimedia Research Group at the IBM T. J. Watson Research Center. He received a B.S. and M.S. degrees in Telecommunications Engineering from the University of Trento, Italy, in 2001 and 2004, respectively, and M.S. and Ph.D. degrees in computer science from Columbia University in 2008 and 2012, respectively. He subsequently joined IBM at the IBM T. J. Watson Research Center, where he has worked on multimedia indexing and retrieval. In 2013, he was co-recipient of an IBM Research Division Award for his work on Multimedia Semantic Modeling. Dr. Merler is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM).

Sharath Pankanti *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (sharat@us.ibm.com).* Dr. Pankanti is a Research Staff Member in the Software Research department at the IBM T. J. Watson Research Center. He received a B.S. degree in electrical and electronics engineering from College of Engineering Pune in 1984, M.Tech. degree in computer science from Hyderabad Central University in 1988, and Ph.D. degree in computer science from the Michigan State University in 1995. He joined IBM at the IBM T. J. Watson Research Center in 1995 as a postdoctoral fellow and in 1996 became Research Staff Member. He is manager of the Exploratory Computer Vision Group at the T. J. Watson Research Center since 2008. He has led a number of safety, productivity, and security focused projects involving biometrics,

multi-sensor surveillance, rail-safety, and driver assistance technologies that entail object/event modeling and detection and recognition from information provided by static and moving sensors/cameras. He is an author or coauthor of more than 20 patents and more than 100 technical papers. Dr. Pankanti is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM).

John R. Smith *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (jsmith@us.ibm.com).* Dr. Smith is a Senior Manager of the Intelligent Information Systems team at the IBM T. J. Watson Research Center. He leads research on multimedia retrieval including image/video content extraction, semantic modeling, video event detection, and social media analysis. Dr. Smith is principal investigator for the IBM Multimedia Analysis and Retrieval System (IMARS). He conducted some of the earliest work on image/video search (VisualSEEk, WebSEEk) and has published more than 200 papers in leading journals and conferences (15.5K citations, h-index of 57, i10-index of 182). Dr. Smith is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE).

Tanveer Syeda-Mahmood *IBM Research - Almaden, San Jose, CA 95120 USA (stf@us.ibm.com).* Dr. Syeda-Mahmood is the Chief Scientist and overall lead for the Medical Sieve Radiology Grand Challenge project in IBM Research - Almaden. She graduated from the MIT AI Lab in 1993 with a Ph.D. degree in computer science. Prior to IBM, she worked as a Research Staff Member at Xerox Webster Research Center, Webster, New York, where she led the image indexing program and was one of the early originators of the field of content-based image and video retrieval. Currently, she is working on applications of content-based retrieval in healthcare and medical imaging. Over the past 30 years, her research interests have been in a variety of areas relating to artificial intelligence including computer vision, image and video databases, medical image analysis, bioinformatics, signal processing, document analysis, and distributed computing frameworks. She has over 200 refereed publications and over 80 filed patents. Dr. Syeda-Mahmood was the General Chair of the First Institute of Electrical and Electronics Engineers (IEEE) International Conference on Healthcare Informatics, Imaging, and Systems Biology in San Jose, California, in 2011. She was also the program co-chair of the IEEE Computer Vision and Pattern Recognition Conference (CVPR 2008). Dr. Syeda-Mahmood is a Fellow of IEEE, a member of IBM Academy of Technology, and an IBM Master Inventor.