Late Fusion of Deep Learning and Handcrafted Visual Features for Biomedical Image Modality Classification

Article *in* IET Image Processing · November 2018

DOI: 10.1049/iet-ipr.2018.5054

Henning Müller
HES-SO Valais-Wallis
760 PUBLICATIONS **10,906** CITATIONS

Late Fusion of Deep Learning and Handcrafted Visual Features for Biomedical Image Modality Classification

Sheng Long Lee [1], Mohammad Reza Zare [2*], Henning Muller [3]

[1] School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Malaysia
[2] School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Malaysia
[3] University of Applied Sciences Western Switzerland (HES-SO) Valais, Sierre, Switzerland
[*] mreza_zare57@yahoo.com

Abstract: Much of medical knowledge is stored in the biomedical literature, collected in archives like PubMed Central that continue to grow rapidly. A significant part of this knowledge is contained in images with limited metadata available which makes it difficult to explore the visual knowledge in the biomedical literature. Thus, extraction of metadata from visual content is important. One important piece of metadata is the type or classification of the image which could be of various medical imaging modalities such as X-ray, computed tomography, or magnetic resonance images. Additionally, they could be general illustrations such as graphs and charts. This paper explores a late score-based fusion of several deep convolutional neural networks with a traditional hand-crafted bag of visual words classifier to classify images from the biomedical literature into image types or modalities. It achieved classification accuracy of 85.51% on the ImageCLEF 2013 modality classification task which is better than the best visual methods and competitive with mixed methods that make use of both visual and textual information. It achieved similarly good results of 84.23% and 87.04% classification accuracy before and after augmentation respectively on the related ImageCLEF 2016 subfigure classification task.

1. Introduction

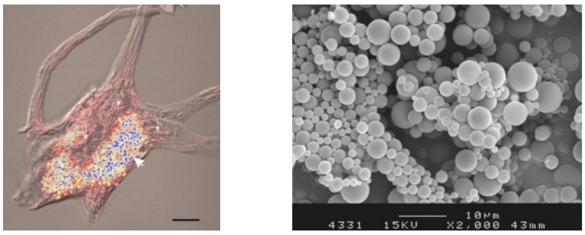
The advent of the internet has inverted many problems of information scarcity to problems of abundance. The sheer volume of information available online has made information retrieval increasingly important in daily life to avoid wasting time on irrelevant information as evidenced by the popularity of search engines such as Google and Bing. The medical field is no exception to this trend with searchable online archives such as PubMed Central which continue to grow rapidly as more articles are submitted and indexed for use by medical practitioners and researchers.

The PubMed Central archive of the biomedical open access literature is a large collection of articles containing text and images that represent an important part of the biomedical knowledge. Visual information plays an important role to represent the knowledge stored but with only little metadata being available it is hard to exploit this information directly. Biomedical image modality classification is the problem of labelling biomedical images with their modality or in a larger sense the image type of the figure.

In medical imaging, a modality is a method or technique to create images such as X-rays, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), or electron microscopy [2]. Biomedical literature also has many other image types such as non-clinical graphs, flowcharts and illustrations. There are also compound figures made up of two or more sub-images each with the same or different modalities. Such compound figures usually need to be separated before classifying the subfigures into their image types [3] but extracting image types without a separation is also possible.

Biomedical image type classification is important to improve medical image retrieval by filtering or reordering results taking into account modality information. Automatic prediction of modality in queries may be used to improve retrieval results [4]. Alternatively, users of medical image retrieval systems have suggested explicit querying by modality would be useful to them [5]. Thus, modality classification is important for retrieval of images lacking explicit modality information as found in the biomedical literature.

Biomedical image type classification is made difficult by low intra-class similarity, imbalanced class distribution and scarce data regarding the wide variety of images in different classes. There are modalities containing images that are visually dissimilar or look different from one another as illustrated in Fig. 1. This low intra-class similarity and the semantic nature of biomedical image modalities make purely unsupervised methods difficult. Supervised methods are challenged by imbalanced class distribution and scarcity of some of the classes in the available datasets as illustrated in Fig. 2.



(a) An electron microscopy image.

(b) Another electron microscopy image.

Fig. 1 *(a), (b) show example images from the ImageCLEF 2013 modality classification sub-task that are visually dissimilar but in the same modality class "DMEL" or electron microscopy.*

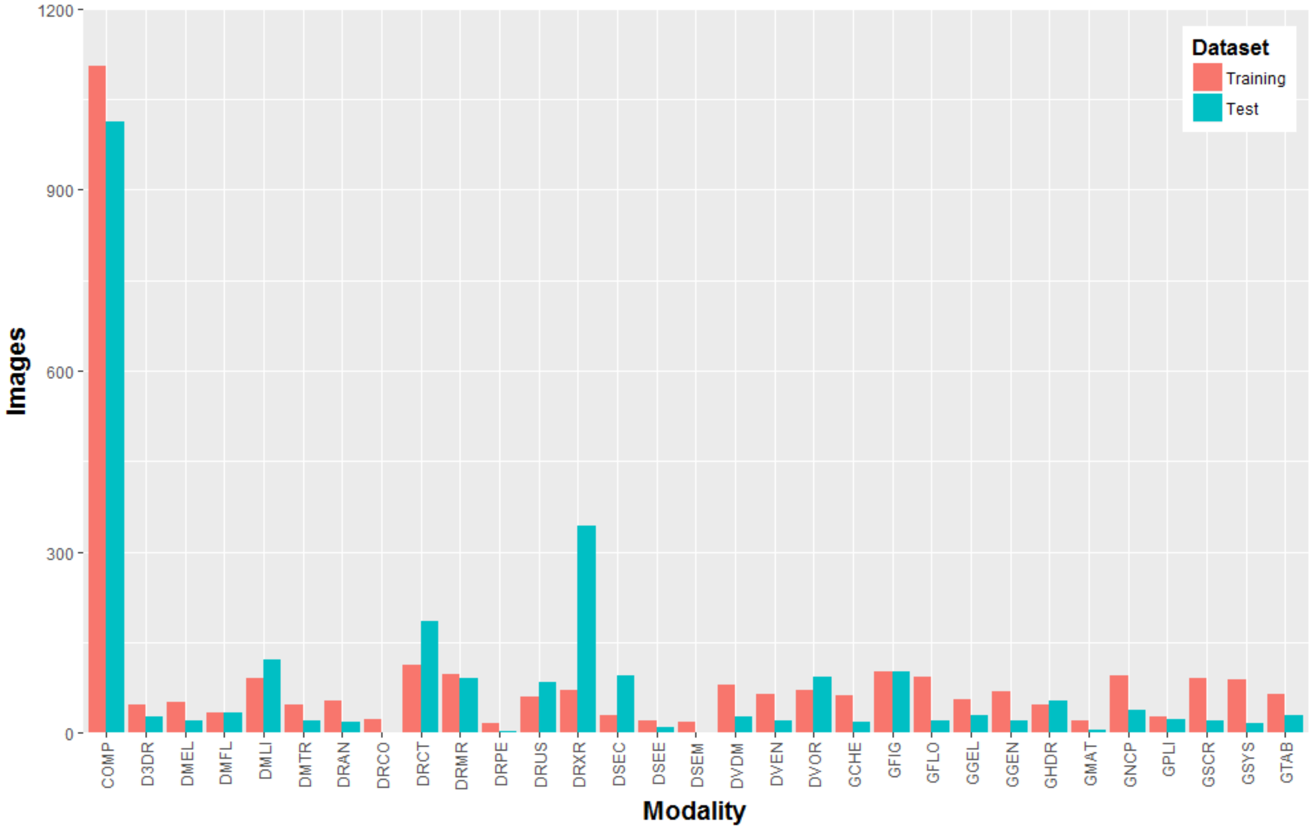


Fig. 2 *The imbalanced class distribution of the ImageCLEF 2013 modality classification dataset that uses a subset of the images in PubMed Central (PMC) [1].*

The classification model proposed in this paper makes use of deep Convolutional Neural Networks (CNNs) for their state-of-the-art performance in other image classification problems [6-10]. Transfer learning was applied to overcome the limited amount of training data that are available and score-based fusion was used for classifier combination of multiple CNNs as well as traditional hand-crafted features to further improve performance. Different CNN architectures, transfer learning methods, and score-based fusion operators were explored. Only CNNs supported by MATLAB without third-party extensions were used.

1.1. Background

The original impetus for most work done in the biomedical image modality classification was the ImageCLEF modality classification or detection sub-task running from 2010 to 2013 [1, 11-13]. The key differences between the datasets are summarized in Table 1.

Table 1 Summary of biomedical image classification datasets over time.

Year	Classes	Training	Test	Total
2010	8	2390	2620	5010
2011	18	1000	1000	2000
2012	31	1000	1000	2000
2013	31	2901	2582	5483
2015	30	4532	2244	6776
2016	30	6776	4166	10942

A similar problem was later reintroduced as the ImageCLEF subfigure classification sub-task in 2015 and 2016 which focused on modality classification of individual subfigures of compound figures, in effect removing the "COMP" or compound figure modality [14, 15].

1.2. Traditional Feature Engineering

Early approaches to biomedical image modality classification adapted methods developed for image processing problems including colour, edge, shape, texture, and transform based image descriptors or visual features [16-23]. The Scale-Invariant Feature Transform (SIFT) [24] was a popular feature choice with good performance [16-19, 23].

The low-level visual features were then used with different classifiers. Support Vector Machines (SVM) were a popular classifier choice [16-19, 22, 23, 25] in biomedical image modality classification and still remain in use [26-30]. Other classifiers such as k-Nearest Neighbours (k-NN) [31], random forests [32], genetic programming [33], and linear regression [34] were attempted but did not perform significantly better than SVMs in many cases.

Different low-level feature selection, combination or fusion criteria were also examined in biomedical image modality classification. Early fusion or feature level fusion using concatenation was the most common followed by late score level fusion using the average score [16-23]. Several papers also found late fusion with rank-based combination to be more stable than score-based fusion. More complex fusion criteria such as Multiple Kernel Learning (MKL) or kernel level fusion [16] and covariance descriptors [35] were also explored.

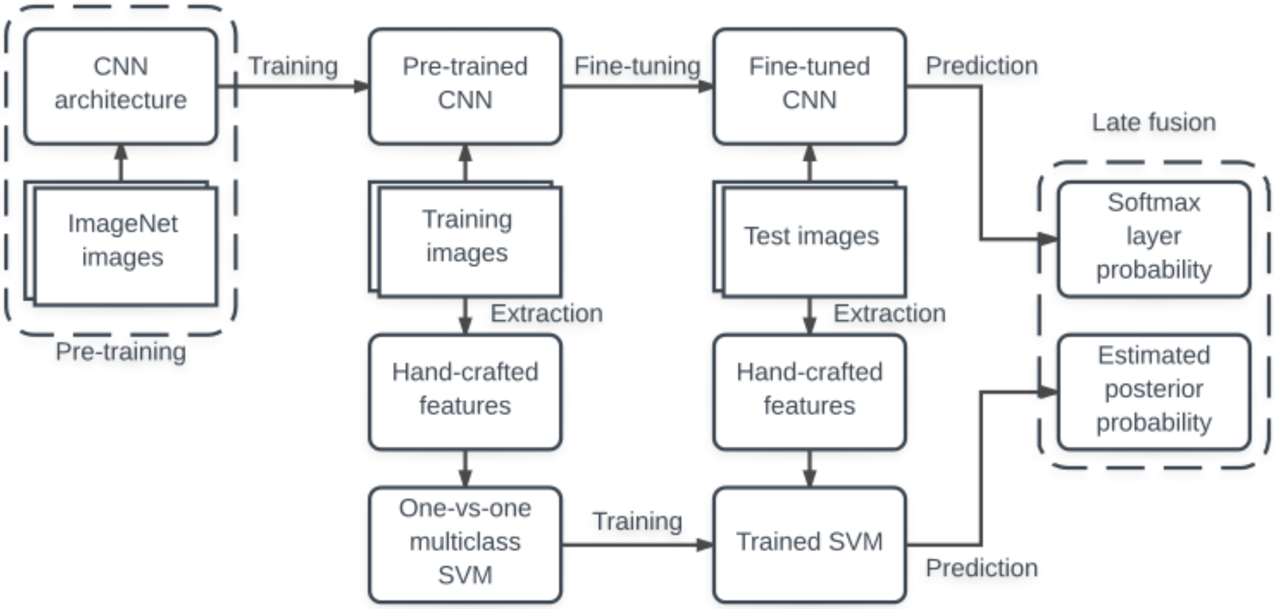


Fig. 3 *Overview of the proposed method for biomedical image modality classification using deep CNNs, transfer learning, hand-crafted features, and late fusion.*

1.3. Deep Representation Learning

Deep learning approaches learn multiple levels of representations or features from input images with each level transforming input into a higher more abstract level until the final output as class labels [36]. Convolutional Neural Networks (CNNs) are the current state-of-the-art deep learning approach for image classification as evidenced by results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [8].

Deep learning approaches require large training datasets to achieve state-of-the-art classification performance, but this requirement can be mitigated using transfer learning. Transfer learning takes representations learned from an extensive training dataset and applies them to a different target dataset or problem which is usually smaller or more limited [37, 38]. There are two main ways to approach transfer learning with CNNs:

1. Extract features from pre-trained CNNs without training CNNs on the target dataset for use in other classifiers like SVMs [39, 40].
2. Fine-tune pre-trained CNNs by adapting last CNN layers and then training on a target dataset to use the trained CNN as classifier [37, 41].

Early deep learning approaches to biomedical image modality classification did not take optimal advantage of transfer learning and were limited by the relatively small datasets [42-44]. Later attempts used transfer learning (mostly with ImageNet), sometimes with data augmentation to expand the available dataset [26, 27, 45]. Fusion was also applied to combinations of different pre-trained CNN architectures to further improve classification performance [29, 45, 46].

2. Methods and Materials

All The datasets used for evaluation were from the ImageCLEF 2013 modality classification and ImageCLEF 2016 subfigure classification sub-tasks, the most recent and extensive datasets for their respective sub-tasks [14, 17]. All datasets are still available for researchers, so reproducibility of the results is given. Additionally, the training set of the "2016" dataset was augmented with all non-compound images from the "2013" dataset for comparison. It is denoted henceforth as "2016augtrn" to distinguish it from the original "2016" dataset without the augmented training set.

The high-level overview of the proposed method is shown in Fig. 3. We use the AlexNet [47], VGG-16, and VGG-19 [48] architectures for transfer learning in the deep learning models due to their image classification performance and compatibility with MATLAB. The pre-trained CNNs were originally trained to compete in ILSVRC on a source dataset of over 1000000 images in 1000 classes [8]. We use a Bag of Visual Words (BoVW) or Bag of Keypoints (BoK) [49] model using SIFT descriptors [24] as a common exemplar for hand-crafted visual features due to their good image classification performance and wide use in biomedical image modality classification [16-19, 23].

2.1. Fine-tuning Pre-trained CNNs for Softmax Classification

The pre-trained CNN is adapted by replacing the last 3 layers specific to the source ImageNet dataset with layers created for the target ImageCLEF dataset. The global learning rate of all the layers in the adapted CNN are lowered whereas the newly created layers are given a multiplier to increase their learning rate. We empirically select a global learning rate of 0.00025 as well as a multiplier of 10 for the newly created layers.

The adapted CNN is then trained using Stochastic Gradient Descent with Momentum (SGDM) in mini-batches of size 32, shuffled every epoch and momentum of 0.9. We use early stopping in addition to the existing dropout and weight decay in the pre-trained CNNs to avoid overfitting. All these parameters were chosen as standard parameters that have shown good performance in other image classification tasks. We hold back 10% per class of the training images as a validation set and train the adapted CNN on the remainder. Training continues until the minimum validation loss evaluated every 50 iterations did not decrease in the past 7 evaluations. Once training stops, the fine-tuned CNN is evaluated on the test images using the built-in softmax layer for classification.

2.2. Hand-crafted Visual Feature Extraction for SVM Classification

The traditional hand-crafted visual feature model is a Bag of Visual Words (BoVW) model adapted from work done by Zare and Müller [50] on a similar dataset for the ImageCLEF 2015 compound figure detection subtask [14]. It was successfully applied to X-ray image classification [51, 52], which is a similar problem and dataset.

A BoVW model begins by detecting or sampling keypoints from the image. We use a Difference of Gaussians (DoG) to detect local interest points and then extract SIFT [24] features around the local interest points or keypoints. Next, the SIFT feature vectors are quantized into K groups or partitions using k-means clustering [49]. Each SIFT feature vector is then assigned to the closest cluster centre using nearest neighbours with Euclidean distance metric. The image is represented by a feature vector or histogram constructed from the frequency that each cluster centre or codeword occurs in the image. The parameter K or codebook size was set to $K = 500$ after empirical testing for a range of values $K = 400, 500, 600$.

However, spatial information is ignored at this stage as all keypoints are assigned equal weight. Spatial information is incorporated into the BoVW model using spatial pyramids [53]. The image is subdivided into L grids such that the level l grid has 2^l cells along each direction for $l = 0, 1, \cdots, L-1$. The codeword frequency histograms are constructed for each cell including the frequencies of the cells that subdivide them. The parameter L or pyramid level was set to $L = 2$ resulting in $2^{0\times2} + 2^{1\times2} = 5$ cells or subdivisions. The final image feature vector produced is of dimension $K\times5 = 2500$.

The adapted model uses the LibSVM library [54] for classification. The Gaussian or Radial Basis Function (RBF) kernel is used in case the features are not linearly separable. The hyperparameter optimization is done using grid search over $C = 2^{-5}, 2^{-3}, \cdots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \cdots, 2^{3}$ to minimize the 10-fold cross validation error on the training images. The range of γ (Gamma) and C (Cost) that determine the impact extent of single training examples and the simplicity of the decision surface respectively are based on recommendations by LibSVM authors [55]. The trained SVM classifier is then evaluated on the feature vectors extracted from the test images and the posterior probabilities are estimated.

2.3. Classifier Combination Using Score-based Late Fusion

Late fusion is quite common in biomedical image modality classification [16-18, 20, 43, 45]. However, usually only the average score is taken which behaves similarly to the "combSUM" fusion operator. We also explore other fusion operators, such as "combMAX", "combMIN" and "combPROD" [56] or "combMED" that takes the median score. The input scores are taken as the estimated posterior probabilities from the SVM and Softmax classifiers.

3. Results and Discussion

The trained individual models are evaluated on the three datasets "2013", "2016", "2016augtrn" and recorded in Table 2. Then, the individual models are combined with score-based late fusion on the same three datasets. However, the possible combinations of individual models are not explored exhaustively but only selected combinations are examined for clarity and conciseness.

3.1. Individual Models

All individual models exhibit a general increase in performance as the size of the dataset increases from 5483 images in "2013" to 10942 images in "2016" and last "2016augtrn" which is the sum of "2013" and "2016" excluding compound images totalling 14306 images. This is expected behaviour for machine learning algorithms.

Deep learning outperforms the traditional hand-crafted feature exemplar of BoVW using SIFT features and spatial pyramids. It seems consistent with the relative performance of deep learning models in other image classification problems. However, it remains possible that there are implementation problems or insufficient optimization due to limited resources.

3.2. Combination Models

The estimated posterior probabilities from SVM and softmax classifiers are used to perform late score-based fusion using the "combSUM", "combPROD", "combMAX", "combMED", and "combMIN" fusion operators. Classifier combinations across different CNN architectures are recorded in Table 3 as it was found to be effective in other similar works [29, 46]. Table 4 investigates effects of combining deep learning and traditional hand-crafted models.

Table 2 Classification performance of each individual model

	Test set accuracy		
Description	2013	2016	2016augtrn
Fine-tuning			
AlexNet	79.01%	80.77%	82.48%
VGG-16	81.10%	**83.63%**	**86.22%**
VGG-19	**83.46%**	82.38%	85.74%
Hand-crafted			
BoVW	66.54%	68.39%	74.24%

Table 3 Classification performance of different CNN architectures combined using score-based fusion operators

	Test set accuracy		
Description	2013	2016	2016augtrn
combMAX	84.47%	84.04%	86.65%
combMED	84.97%	**84.97%**	86.99%
combMIN	84.28%	84.16%	86.89%
combPROD	**85.09%**	84.78%	87.30%
combSUM	85.05%	84.78%	**87.35%**

Table 4 Classification performance of different CNN architectures and hand-crafted visual features combined using score-based fusion operators

	Test set accuracy		
Description	2013	2016	2016augtrn
combMAX	84.51%	83.56%	86.41%
combMED	85.44%	**84.52%**	87.04%
combMIN	84.39%	82.57%	86.94%
combPROD	**85.52%**	84.23%	87.04%
combSUM	85.36%	84.16%	**87.18%**

Fig. 4 compares the classification performance of different score-based fusion operators. The aggregation based fusion operators “combPROD” and “combSUM” seem to outperform the selection based fusion operators “combMAX” and “combMIN” but “combMED” is at a similar level.

Fig. 4 also compares classification performance of combinations with and without hand-crafted visual features. Combinations with hand-crafted visual features have better performance on the “2013” dataset but have worse performance on the “2016” dataset. It also affects the “2016augtrn” dataset as most images are from the “2016” dataset resulting in slightly worse performance.

Table 5 Classification performance baselines of other models for comparison

	Test set accuracy		
Description	2013	2016	2016augtrn
Visual only models			
2013 best visual [16]	80.79%	-	-
2016 best visual [26]	-	-	85.38%
Yu *et al.* [45]	-	82.61%	**87.37%**
Kumar *et al.* [29]	-	82.48%	-
Valavanis *et al.* [30]	83.04%	82.45%	85.19%
Our proposed model	**85.51%**	**84.23%**	87.04%
Mixed models (visual and textual)			
2013 best mixed [16]	81.68%	-	-
2016 best mixed [26]	-	-	**88.43%**
Valavanis *et al.*[30]	**85.71%**	**86.10%**	88.07%

3.3. *Proposed Model*

The combination with the best test set accuracy for the “2013” dataset is a combination of fine-tuned AlexNet, VGG-16, VGG-19, and hand-crafted BoVW models using the “combPROD” fusion operator. However, it does not obtain the best classification performance in the “2016” dataset as previously noted. We prioritize the “2013” dataset, as it more closely resembles the distribution of images in biomedical literature by including compound images [1].

Table 5 compares the accuracy of our proposed model with other models as baselines. We achieve good performance compared with other visual methods using only input images. Mixed methods utilizing text input such as captions in addition to input images still perform better. The models with best classification performance for each category and dataset are highlighted in bold.

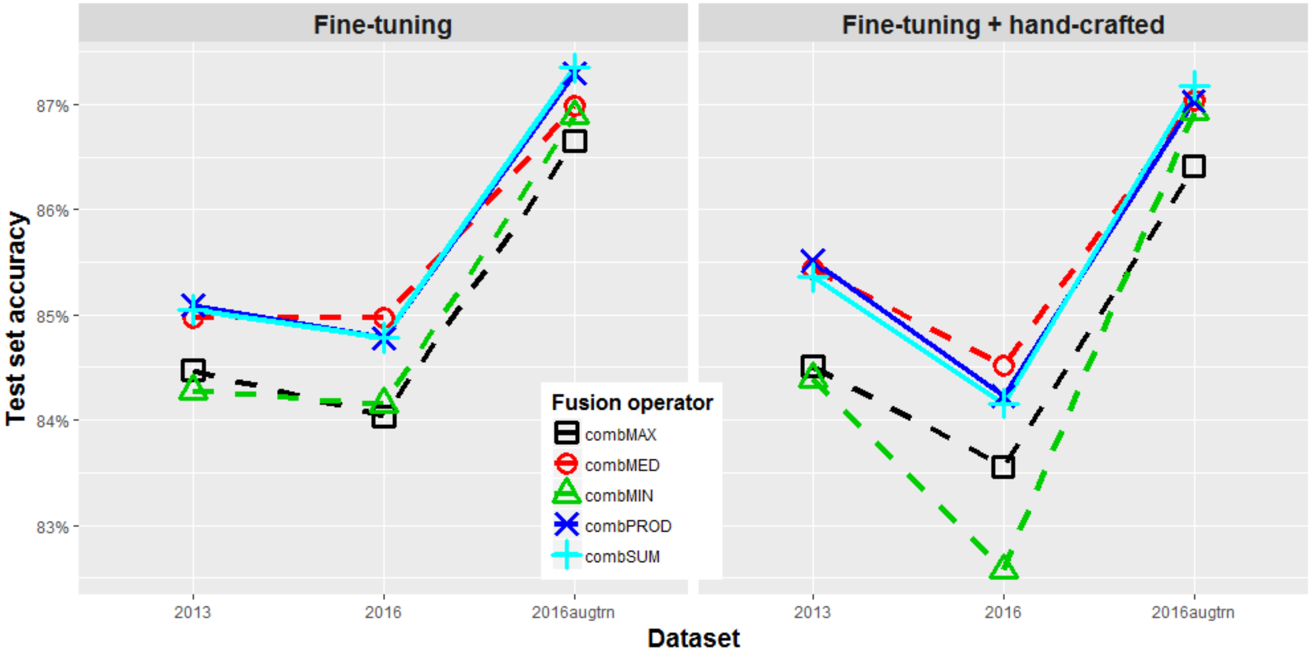


Fig. 4 Classification performance of fusion models over different datasets combining different CNN architectures with and without hand-crafted visual features.

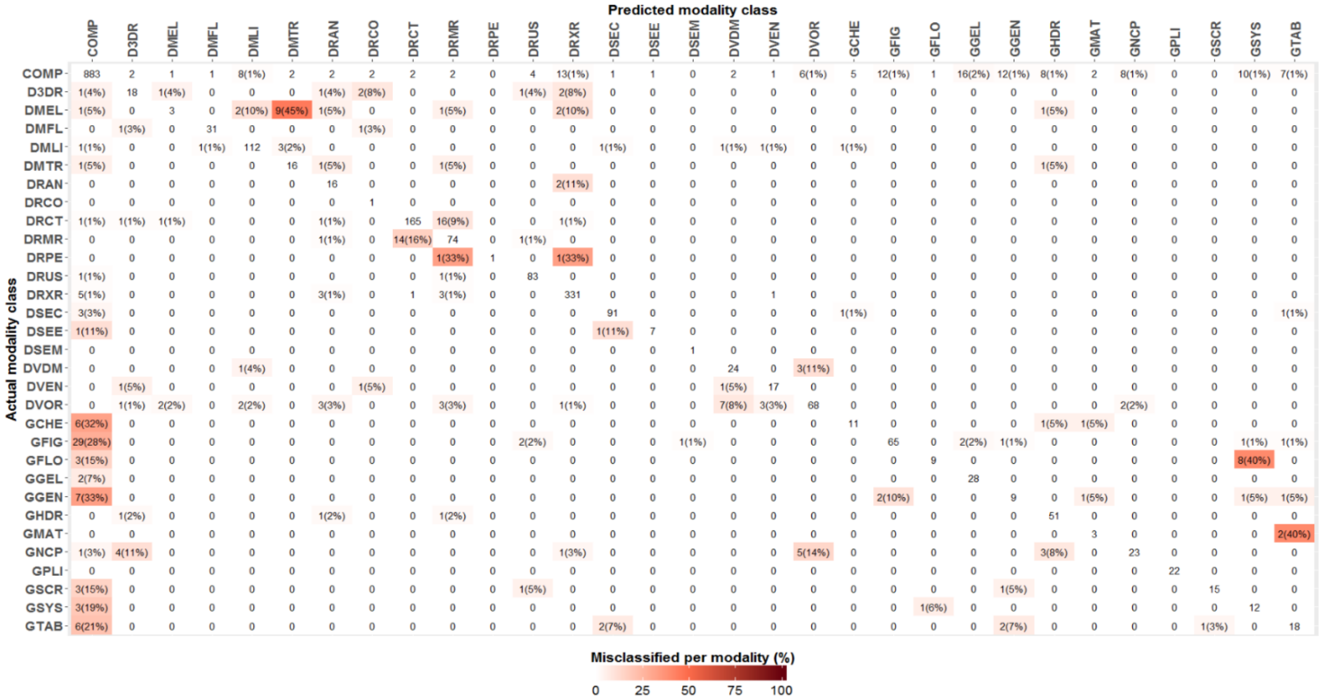


Fig. 5 *Confusion matrix of the proposed model on the "2013" test images.*

3.4. Classifier Analysis

After selecting the proposed model, we examine the classification results in detail to see what was classified correctly and what was misclassified. Table 6 in the appendix records the detailed classification results for each modality or class for the 2013, 2016, and 2016augtrn datasets. It presents the precision, recall and F1 score or F-measure which are preferred performance measures for imbalanced datasets.

Fig. 5 shows the most common misclassifications by the proposed model on the "2013" dataset. It is dominated by the "COMP" or compound figure class. Compound images are misclassified as many other classes because they are composed of subfigures of other modality classes. Inversely, generic biomedical illustrations such as "GFIG" are misclassified as compound images. It is likely due to whitespace in the illustrations being mistaken for subfigure separation in compound figures. "DRCT" and "DRMR", also known as CT and MRI images, are also commonly misclassified as one another.

"COMP" or compound figures are no longer a source of misclassifications in the "2016" and "2016augtrn" datasets. Instead, Fig. 6 shows many classes are misclassified

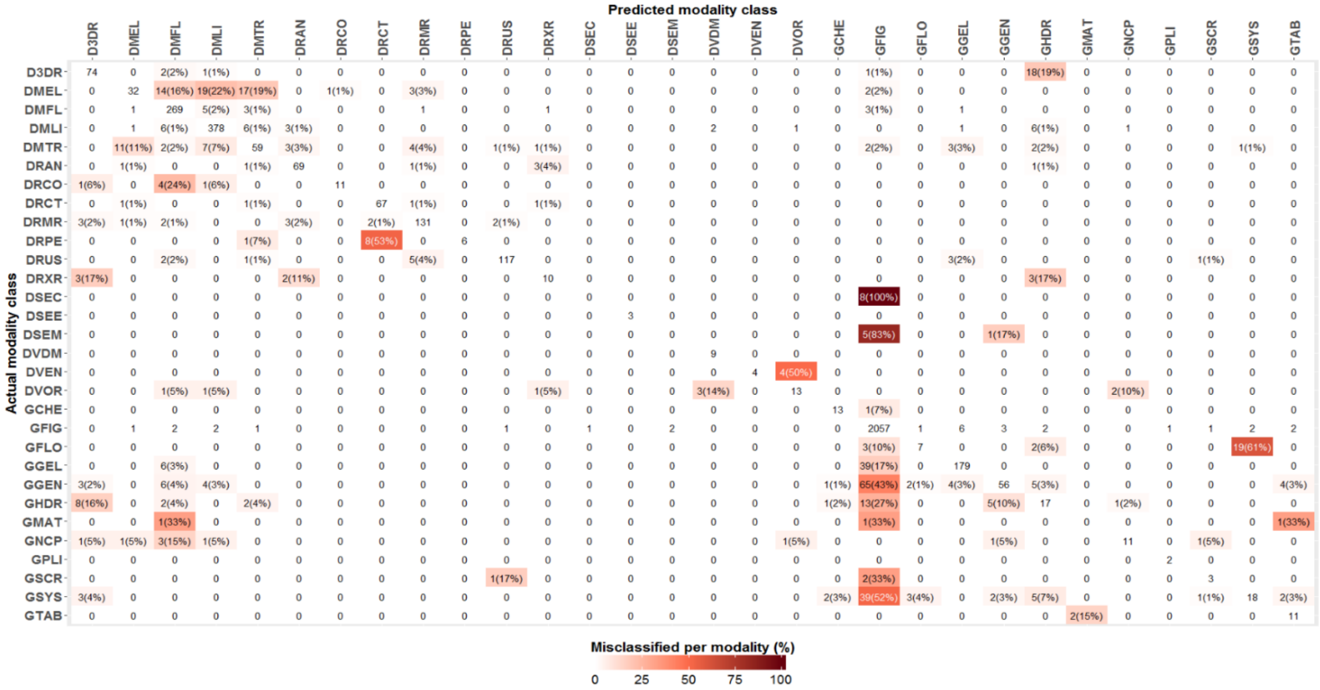
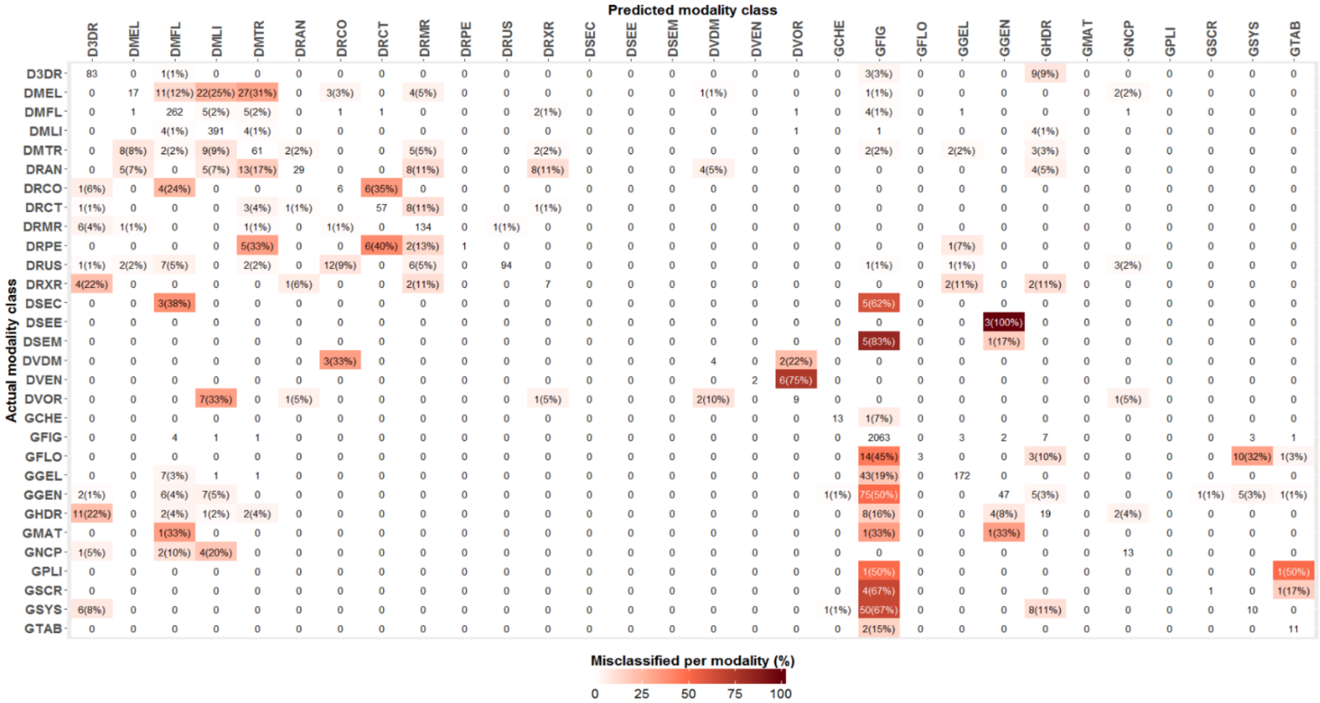


Fig. 6 *Confusion matrix of the proposed model on the "2016" test images.*



	D3DR	DMEL	DMFL	DMLI	DMTR	DRAN	DRCO	DRCT	DRMR	DRPE	DRUS	DRXR	DSEC	DSEE	DSEM	DVDM	DVEN	DVOR	GCHE	GFIG	GFLO	GGEL	GGEN	GHDR	GMAT	GNCP	GPLI	GSCR	GSYS	GTAB
D3DR	83	0	1(1%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3(3%)	0	0	0	9(9%)	0	0	0	0	0	0
DMEL	0	17	11(12%)	22(25%)	27(31%)	0	3(3%)	0	4(5%)	0	0	0	0	0	0	1(1%)	0	0	0	1(1%)	0	0	0	0	0	2(2%)	0	0	0	0
DMFL	0	1	262	5(2%)	5(2%)	0	1	1	0	0	0	2(1%)	0	0	0	0	0	1	0	4(1%)	0	1	0	0	0	1	0	0	0	0
DMLI	0	0	4(1%)	391	4(1%)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	4(1%)	0	0	0	0	0	0
DMTR	0	8(8%)	2(2%)	9(9%)	61	2(2%)	0	0	5(5%)	0	0	2(2%)	0	0	0	0	0	0	0	2(2%)	0	2(2%)	0	3(3%)	0	0	0	0	0	0
DRAN	0	5(7%)	0	5(7%)	13(17%)	29	0	0	8(11%)	0	0	8(11%)	0	0	0	4(5%)	0	0	0	0	0	0	0	4(5%)	0	0	0	0	0	0
DRCO	1(6%)	0	4(24%)	0	0	0	6	6(35%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DRCT	1(1%)	0	0	0	3(4%)	1(1%)	0	57	8(11%)	0	0	1(1%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DRMR	6(4%)	1(1%)	0	0	1(1%)	0	1(1%)	0	134	0	1(1%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DRPE	0	0	0	0	5(33%)	0	0	6(40%)	2(13%)	1	0	0	0	0	0	0	0	0	0	0	0	1(7%)	0	0	0	0	0	0	0	0
DRUS	1(1%)	2(2%)	7(5%)	0	2(2%)	0	12(9%)	0	6(5%)	0	94	0	0	0	0	0	0	0	0	1(1%)	0	1(1%)	0	0	0	3(2%)	0	0	0	0
DRXR	4(22%)	0	0	0	0	1(6%)	0	0	2(11%)	0	0	7	0	0	0	0	0	0	0	0	0	2(11%)	0	2(11%)	0	0	0	0	0	0
DSEC	0	0	3(38%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5(62%)	0	0	0	0	0	0	0	0	0	0
DSEE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3(100%)	0	0	0	0	0	0	0
DSEM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5(83%)	0	0	1(17%)	0	0	0	0	0	0	0
DVDM	0	0	0	0	0	0	3(33%)	0	0	0	0	0	0	0	0	4	0	2(22%)	0	0	0	0	0	0	0	0	0	0	0	0
DVEN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6(75%)	0	0	0	0	0	0	0	0	0	0	0	0
DVOR	0	0	0	7(33%)	0	1(5%)	0	0	0	0	0	1(5%)	0	0	0	2(10%)	0	9	0	0	0	0	0	0	0	1(5%)	0	0	0	0
GCHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	1(7%)	0	0	0	0	0	0	0	0	0	0
GFIG	0	0	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2063	0	3	2	7	0	0	0	0	3	1
GFLO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14(45%)	3	0	0	3(10%)	0	0	0	0	10(32%)	1(3%)
GGEL	0	0	7(3%)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43(19%)	0	172	0	0	0	0	0	0	0	0
GGEN	2(1%)	0	6(4%)	7(5%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1(1%)	75(50%)	0	0	47	5(3%)	0	0	0	1(1%)	5(3%)	1(1%)
GHDR	11(22%)	0	2(4%)	1(2%)	2(4%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8(16%)	0	0	4(8%)	19	0	2(4%)	0	0	0	0
GMAT	0	0	1(33%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1(33%)	0	0	1(33%)	0	0	0	0	0	0	0
GNCP	1(5%)	0	2(10%)	4(20%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0
GPLI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1(50%)	0	0	0	0	0	0	0	0	0	1(50%)
GSCR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4(67%)	0	0	0	0	0	0	0	1	0	1(17%)
GSYS	6(8%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1(1%)	50(67%)	0	0	0	8(11%)	0	0	0	0	10	0
GTAB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2(15%)	0	0	0	0	0	0	0	0	0	11

***Fig.** 7 Confusion matrix of the proposed model on the "2016augtrn" test images*

as the new dominant class "GFIG" or statistical figures, graphs, and charts. Microscopy images with modality code beginning with "DM" are often misclassified as "DMEL" or electron microscopy due to the visual dissimilarity within the class as illustrated in Fig. 1. Fig. 7 shows the same patterns as Fig. 7 with a smaller proportion of misclassified images due to the augmented or expanded dataset.

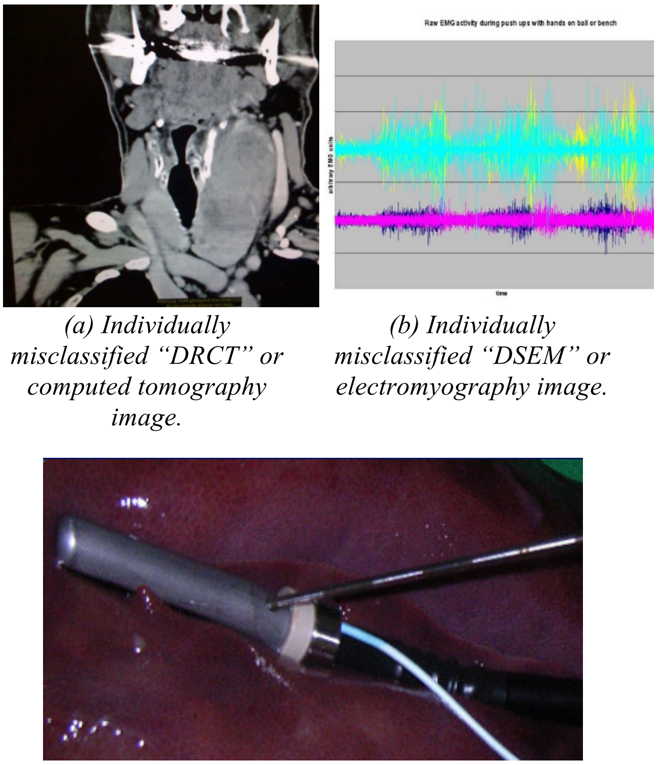
3.5. Combination Analysis

After examining the proposed model in detail, we examine how each individual classification model in the late fusion contributes to the proposed model. Table 7 in the appendix records the proportion of each modality class that was misclassified for each individual model as well as the proposed late fusion model evaluated on the "2013" dataset. The best or lowest proportion misclassified are highlighted in bold per modality class.

Table 7 shows that each individual model has classes or modalities it is best at with lowest proportion of misclassified images. It suggests that the individual models differ in which modality they are good at classifying. Hence, the individual models are good candidates for combination with late fusion.

From the misclassification of the proposed model, it can be observed that there are modalities where the proposed model has a lower misclassification rate than even the best individual model misclassification rate. The decrease in the misclassification rate is mostly due to the late fusion selecting the best prediction from the individual model predictions.

However, there are also images that the proposed late fusion model classified correctly that were misclassified in all the individual models as illustrated in Fig. 8. It suggests that there is a synergistic effect from the late fusion of the individual models.


Raw EMG activity during push ups with hands on ball or bench

(a) Individually misclassified "DRCT" or computed tomography image.

(b) Individually misclassified "DSEM" or electromyography image.

(c) Individually misclassified "DVOR" or visible light photography of other organs.

***Fig. 8** (a), (b), (c) A few example images from the ImageCLEF 2013 modality classification sub-task that were classified correctly by the proposed late fusion model but misclassified by individual models in the combination.*

4. Conclusion

A late fusion model was proposed that combined deep learning models and traditional hand-crafted visual features for biomedical image modality classification. Transfer learning was used to mitigate the limited and imbalanced dataset. Specifically, fine-tuning of pre-trained CNNs AlexNet, VGG-16, and VGG-19 with early stopping was found to be effective relative to CNN feature extraction into SVM classifiers.

A traditional hand-crafted BoVW model using SIFT features was found to improve performance of the combined classifier although it had worse individual performance. A relatively simple late fusion method with the score-based fusion operator "combPROD" was found to be sufficient compared to more complex methods like stacked SVMs.

The proposed model outperforms or is similar to other visual methods on two separate but similar datasets, the ImageCLEF 2013 modality classification sub-task and the ImageCLEF 2016 subfigure classification sub-task as shown in Table 5. However, it still falls short of mixed methods that use both visual and text input.

Future work or improvements may include combining more CNN architectures or even training from scratch, but this also increases the computational time required and hence decreases efficiency. Other traditional hand-crafted features could be included in the combination as well. More complex combination schemes such as Multiple Kernel Learning (MKL) or kernel level fusion could be evaluated. Finally, mixed methods incorporating textual input such as image captions can be explored.

5. References

1. Garcia Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., and Müller, H., 'Overview of the Imageclef 2013 Medical Task', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)
2. Bushberg, J.T., Seibert, J.A., Leidholdt, E.M., and Boone, J.M., *The Essential Physics of Medical Imaging, Third Edition*, (Lippincott Williams & Wilkins, 2011)
3. Chhatkuli, A., Foncubierta-Rodríguez, A., Markonis, D., Meriaudeau, F., and Müller, H., 'Separating Compound Figures in Journal Articles to Allow for Subfigure Classification', in, *SPIE Medical Imaging 2013: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, (International Society for Optics and Photonics, 2013)
4. Tirilly, P., Lu, K., Mu, X., Zhao, T., and Cao, Y., On Modality Classification and Its Use in Text-Based Image Retrieval in Medical Databases', *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, (IEEE, 2011)
5. Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., and Müller, H., 'A Survey on Visual Information Search Behavior and Requirements of Radiologists', *Methods of Information in Medicine*, 2012, 51, (6), pp. 539-548.
6. Banerjee, I., Crawley, A., Bhethanabotla, M., Daldrup-Link, H.E., and Rubin, D.L., 'Transfer Learning on Fused Multiparametric Mr Images for Classifying Histopathological Subtypes of Rhabdomyosarcoma', *Computerized Medical Imaging and Graphics*, 2017.
7. Li, H., Mao, Y., Yin, Z., and Xu, Y., 'A Hierarchical Convolutional Neural Network for Vesicle Fusion Event Classification', *Computerized Medical Imaging and Graphics*, 2017, 60, pp. 22-34.
8. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L., 'Imagenet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision*, 2015, 115, (3), pp. 211-252.
9. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O., and Hufnagl, P., 'Deep Convolutional Neural Networks for Automatic Classification of Gastric Carcinoma Using Whole Slide Images in Digital Histopathology', *Computerized Medical Imaging and Graphics*, 2017.
10. Sun, W., Tseng, T.-L., Zhang, J., and Qian, W., 'Enhancing Deep Convolutional Neural Network Scheme for Breast Cancer Diagnosis with Unlabeled Data', *Computerized Medical Imaging and Graphics*, 2017, 57, pp. 4-9.
11. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Reisetter, J., E. Kahn Jr., C., and Hersh, W., 'Overview of the Clef 2010 Medical Image Retrieval Track', in, *CLEF 2010 (Cross Language Evaluation Forum)*, (2010)
12. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., and Tsikrika, T., 'Overview of the Clef 2011 Medical Image Classification and Retrieval Tasks', in, *CLEF 2011 (Cross Language Evaluation Forum)*, (2011)
13. Müller, H., de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.K., and Eggel, I., 'Overview of the Imageclef 2012 Medical Image Retrieval and Classification Tasks', in, *CLEF 2012 (Cross Language Evaluation Forum)*, (2012)
14. Garcia Seco de Herrera, A., Bromuri, S., and Müller, H., 'Overview of the Imageclef 2015 Medical Task', in, *CLEF 2015 (Cross Language Evaluation Forum)*, (2015)
15. Garcia Seco de Herrera, A., Schaer, R., Bromuri, S., and Müller, H., 'Overview of the Imageclef 2016 Medical Task', in, *CLEF 2016 (Cross Language Evaluation Forum)*, (2016)
16. Abedini, M., Cao, L., Codella, N., Connell, J.H., Garnavi, R., Geva, A., Merler, M., Nguyen, Q.-B., Pankanti, S.U., and Smith, J.R., 'Ibm at Imageclef 2013 Medical Tasks', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

17. Dimitrovski, I., Kocev, D., Kitanovski, I., Loskovska, S., and Džeroski, S., 'Improved Medical Image Modality Classification Using a Combination of Visual and Textual Features', *Computerized Medical Imaging and Graphics*, 2015, 39, pp. 14-26.

18. Garcia Seco de Herrera, A., Markonis, D., Schaer, R., Eggel, I., and Müller, H., 'The Medgift Group in Imageclefmed 2013', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

19. Kitanovski, I., Dimitrovski, I., and Loskovska, S., 'Fcse at Medical Tasks of Imageclef 2013', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

20. Mourão, A., Martins, F., and Magalhães, J., 'Novasearch on Medical Imageclef 2013', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

21. Simpson, M.S., You, D., Rahman, M.M., Demner-Fushman, D., Antani, S., and Thoma, G., 'Iti's Participation in the 2013 Medical Track of Imageclef', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

22. Simpson, M.S., You, D., Rahman, M.M., Xue, Z., Demner-Fushman, D., Antani, S., and Thoma, G., 'Literature-Based Biomedical Image Classification and Retrieval', *Computerized Medical Imaging and Graphics*, 2015, 39, pp. 3-13.

23. Zhou, X., Han, M., Song, Y., and Li, Q., 'Fast Filtering Techniques in Medical Image Classification and Retrieval', in, *CLEF 2013 (Cross Language Evaluation Forum)*, (2013)

24. Lowe, D.G., Object Recognition from Local Scale-Invariant Features', *Proceedings of the Seventh IEEE International Conference on Computer Vision*, (IEEE, 1999)

25. Stathopoulos, S. and Kalamboukis, T., 'Applying Latent Semantic Analysis to Large-Scale Medical Image Databases', *Computerized Medical Imaging and Graphics*, 2015, 39, pp. 27-34.

26. Koitka, S. and Friedrich, C.M., 'Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of Imageclef 2016 Fhdo Biomedical Computer Science Group (Bcsg)', in, *CLEF 2016 (Cross Language Evaluation Forum)*, (2016)

27. Kumar, A., Lyndon, D., Kim, J., and Feng, D., 'Subfigure and Multi-Label Classification Using a Fine-Tuned Convolutional Neural Network', in, *CLEF 2016 (Cross Language Evaluation Forum)*, (2016)

28. Valavanis, L., Stathopoulos, S., and Kalamboukis, T., 'Ipl at Clef 2016 Medical Task', in, *CLEF 2016 (Cross Language Evaluation Forum)*, (2016)

29. Kumar, A., Kim, J., Lyndon, D., Fulham, M., and Feng, D., 'An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification', *IEEE journal of biomedical and health informatics*, 2017, 21, (1), pp. 31-40.

30. Valavanis, L., Stathopoulos, S., and Kalamboukis, T., Fusion of Bag-of-Words Models for Image Classification in the Medical Domain', in Jose, J.M., Hauff, C., Altıngovde, I.S., Song, D., Albakour, D., Watt, S., and Tait, J. (eds.), *Advances in Information Retrieval: 39th European Conference on Ir Research, Ecir 2017, Aberdeen, Uk, April 8-13, 2017, Proceedings*, (Springer International Publishing, 2017)

31. Markonis, D., Eggel, I., de Herrera, A.G.S., and Müller, H., 'The Medgift Group in Imageclefmed 2011', in, *CLEF 2011 (Cross Language Evaluation Forum)*, (2011)

32. Marée, R., Stern, O., and Geurts, P., 'Biomedical Imaging Modality Classification Using Bags of Visual and Textual Terms with Extremely Randomized Trees: Report of Imageclef 2010 Experiments', in, *CLEF 2010 (Cross Language Evaluation Forum)*, (2010)

33. Faria, F.A., Calumby, R.T., and Torres, R.d.S., 'Recod at Imageclef 2011: Medical Modality Classification Using Genetic Programming', in, *CLEF 2011 (Cross Language Evaluation Forum)*, (2011)

34. Castellanos, A., Benavent, J., Benavent, X., García-Serrano, A., and Ves, E.d., 'Using Visual Concept Features in a Multimodal Retrieval System for the Medical Collection at Imageclef2012', in, *CLEF 2012 (Cross Language Evaluation Forum)*, (2012)

35. Cirujeda, P. and Binefa, X., 'Medical Image Classification Via 2d Color Feature Based Covariance Descriptors', in, *CLEF 2015 (Cross Language Evaluation Forum)*, (2015)

36. LeCun, Y., Bengio, Y., and Hinton, G., 'Deep Learning', *Nature*, 2015, 521, (7553), pp. 436-444.

37. Oquab, M., Bottou, L., Laptev, I., and Sivic, J., Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2014)

38. Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R.M., 'Deep Convolutional Neural Networks for Computer-Aided Detection: Cnn Architectures, Dataset Characteristics and Transfer Learning', *IEEE Transactions on Medical Imaging*, 2016, 35, (5), pp. 1285-1298.

39. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T., Decaf: A Deep Convolutional Activation Feature for Generic Visual Recognition', (PMLR, 2014)

40. Razavian, A.S., Azizpour, H., Sullivan, J., and Carlsson, S., Cnn Features Off-the-Shelf: An Astounding

Baseline for Recognition', *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (IEEE, 2014)

41. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., and Liang, J., 'Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?', *IEEE Transactions on Medical Imaging*, 2016, 35, (5), pp. 1299-1312.

42. Lyndon, D., Kumar, A., Kim, J., Leong, P.H.W., and Feng, D., 'Convolutional Neural Networks for Subfigure Classification', in, *CLEF 2015 (Cross Language Evaluation Forum)*, (2015)

43. Yu, Y., Lin, H., Yu, Q., Meng, J., Zhao, Z., Li, Y., and Zuo, L., 'Modality Classification for Medical Images Using Multiple Deep Convolutional Neural Networks', *Journal of Computational Information Systems*, 2015, 11, pp. 5403-5413.

44. Semedo, D. and Magalhães, J., 'Novasearch at Imageclefmed 2016 Subfigure Classification Task', in, *CLEF 2016 (Cross Language Evaluation Forum)*, (2016)

45. Yu, Y., Lin, H., Meng, J., Wei, X., Guo, H., and Zhao, Z., 'Deep Transfer Learning for Modality Classification of Medical Images', *Information*, 2017, 8, (3), p. 91.

46. Lee, S.L. and Zare, M.R., 'Biomedical Compound Figure Detection Using Deep Learning and Fusion Techniques', *IET Image Processing*, 2018, 12, (6), pp. 1031-1037.

47. Krizhevsky, A., Sutskever, I., and Hinton, G.E., 'Imagenet Classification with Deep Convolutional Neural Networks', in, *Advances in neural information processing systems*, (2012)

48. Simonyan, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition', *3rd International Conference on Learning Representations*, (2015)

49. Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C., 'Visual Categorization with Bags of Keypoints', in, *Workshop on statistical learning in computer vision, ECCV*, (Prague, 2004)

50. Zare, M.R. and Müller, H., 'A Medical X-Ray Image Classification and Retrieval System', in, *PACIS*, (2016)

51. Zare, M.R., Mueen, A., Awedh, M., and Seng, W.C., Automatic Classification of Medical X-Ray Images: Hybrid Generative-Discriminative Approach', *IET Image Processing*, (Institution of Engineering and Technology, 2013)

52. Zare, M.R., Mueen, A., and Seng, W.C., 'Automatic Classification of Medical X-Ray Images Using a Bag of Visual Words', *IET Computer Vision*, 2013, 7, (2), pp. 105-114.

53. Lazebnik, S., Schmid, C., and Ponce, J., Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories', *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, (IEEE)

54. Chang, C.-C. and Lin, C.-J., 'Libsvm', *ACM Transactions on Intelligent Systems and Technology*, 2011, 2, (3), pp. 1-27.

55. Hsu, C.-W., Chang, C.-C., and Lin, C.-J., 'A Practical Guide to Support Vector Classification', 2003.

56. García Seco de Herrera, A. and Müller, H., Fusion Techniques in Biomedical Information Retrieval', *Fusion in Computer Vision*, (Springer International Publishing, 2014)

6. Appendices

Table 6 records detailed classification results for each modality or class for the “2013”, “2016”, and “2016augtrn” datasets. It presents the precision, recall and F1 score or F-measure which are preferred performance measures for imbalanced datasets. It is provided for detailed comparison with other classification results.

Table 7 records the proportion of each modality class that was misclassified for each individual model as well as the proposed late fusion model evaluated on the “2013” dataset. The best or lowest proportion misclassified are highlighted in bold per modality class. It is provided to show the contribution of each individual model to the classification performance of the combined model.

Table 6 Detailed classification results per modality class over different datasets

Modality	2013			2016			2016augtrn		
class	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
COMP	0.92	0.87	0.90	-	-	-	-	-	-
D3DR	0.62	0.69	0.65	0.72	0.86	0.78	0.77	0.77	0.77
DMEL	0.38	0.15	0.21	0.50	0.19	0.28	0.64	0.36	0.46
DMFL	0.94	0.94	0.94	0.83	0.92	0.87	0.84	0.95	0.89
DMLI	0.90	0.93	0.91	0.86	0.97	0.91	0.90	0.93	0.92
DMTR	0.53	0.80	0.64	0.49	0.64	0.55	0.64	0.61	0.63
DRAN	0.53	0.89	0.67	0.85	0.38	0.53	0.86	0.91	0.88
DRCO	0.14	1.00	0.25	0.23	0.35	0.28	0.92	0.65	0.76
DRCT	0.91	0.89	0.90	0.81	0.80	0.81	0.87	0.94	0.91
DRMR	0.72	0.82	0.77	0.79	0.93	0.86	0.90	0.91	0.90
DRPE	1.00	0.33	0.50	1.00	0.07	0.13	1.00	0.40	0.57
DRUS	0.90	0.98	0.94	0.99	0.73	0.84	0.96	0.91	0.93
DRXR	0.94	0.96	0.95	0.33	0.39	0.36	0.59	0.56	0.57
DSEC	0.95	0.95	0.95	0.00	0.00	0.00	0.00	0.00	0.00
DSEE	0.88	0.78	0.82	0.00	0.00	0.00	1.00	1.00	1.00
DSEM	0.50	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
DVDM	0.69	0.86	0.76	0.36	0.44	0.40	0.64	1.00	0.78
DVEN	0.74	0.85	0.79	1.00	0.25	0.40	1.00	0.50	0.67
DVOR	0.83	0.74	0.78	0.47	0.43	0.45	0.68	0.62	0.65
GCHE	0.61	0.58	0.59	0.87	0.93	0.90	0.76	0.93	0.84
GFIG	0.82	0.64	0.72	0.90	0.99	0.94	0.92	0.99	0.95
GFLO	0.82	0.45	0.58	1.00	0.10	0.18	0.54	0.23	0.32
GGEL	0.61	0.93	0.74	0.95	0.77	0.85	0.91	0.80	0.85
GGEN	0.36	0.43	0.39	0.81	0.31	0.45	0.82	0.37	0.51
GHDR	0.78	0.94	0.86	0.30	0.39	0.34	0.28	0.35	0.31
GMAT	0.43	0.60	0.50	0.00	0.00	0.00	0.00	0.00	0.00
GNCP	0.70	0.62	0.66	0.59	0.65	0.62	0.73	0.55	0.63
GPLI	1.00	1.00	1.00	0.00	0.00	0.00	0.67	1.00	0.80
GSCR	0.94	0.75	0.83	0.50	0.17	0.25	0.43	0.50	0.46
GSYS	0.38	0.75	0.50	0.36	0.13	0.19	0.45	0.24	0.31
GTAB	0.60	0.62	0.61	0.69	0.85	0.76	0.55	0.85	0.67

Table 7 Percentage per modality class misclassified of each individual model and proposed combined model

	Misclassified per modality				
Modality	Fine-tuning			Hand-crafted	Proposed
class	AlexNet	VGG-16	VGG-19	BoVW	combined model
COMP	15.78%	18.24%	13.91%	14.30%	**12.92%**
D3DR	53.85%	**30.77%**	**30.77%**	50.00%	**30.77%**
DMEL	95.00%	90.00%	**80.00%**	85.00%	85.00%
DMFL	**6.06%**	9.09%	9.09%	54.55%	**6.06%**
DMLI	19.83%	13.22%	**6.61%**	33.88%	7.44%
DMTR	35.00%	**20.00%**	30.00%	80.00%	**20.00%**
DRAN	22.22%	**11.11%**	16.67%	44.44%	**11.11%**
DRCO	100.00%	**0.00%**	100.00%	100.00%	**0.00%**
DRCT	15.05%	16.67%	12.90%	25.27%	**11.29%**
DRMR	21.11%	20.00%	20.00%	36.67%	**17.78%**
DRPE	66.67%	66.67%	66.67%	**33.33%**	66.67%
DRUS	11.76%	5.88%	4.71%	24.71%	**2.35%**
DRXR	8.72%	5.81%	5.52%	31.40%	**3.78%**
DSEC	19.79%	**5.21%**	15.63%	22.92%	**5.21%**
DSEE	22.22%	22.22%	22.22%	22.22%	22.22%
DSEM	100.00%	100.00%	100.00%	100.00%	**0.00%**
DVDM	50.00%	25.00%	**10.71%**	75.00%	14.29%
DVEN	30.00%	30.00%	**5.00%**	50.00%	15.00%
DVOR	27.17%	38.04%	29.35%	65.22%	**26.09%**
GCHE	**42.11%**	**42.11%**	**42.11%**	52.63%	**42.11%**
GFIG	47.06%	47.06%	**36.27%**	94.12%	**36.27%**
GFLO	55.00%	80.00%	**45.00%**	55.00%	55.00%
GGEL	16.67%	**6.67%**	**6.67%**	53.33%	**6.67%**
GGEN	57.14%	**38.10%**	61.90%	80.95%	57.14%
GHDR	16.67%	**3.70%**	12.96%	83.33%	5.56%
GMAT	60.00%	**40.00%**	80.00%	80.00%	**40.00%**
GNCP	51.35%	**32.43%**	51.35%	70.27%	37.84%
GPLI	18.18%	18.18%	**9.09%**	13.64%	**0.00%**
GSCR	45.00%	**25.00%**	**25.00%**	80.00%	**25.00%**
GSYS	50.00%	**12.50%**	37.50%	100.00%	25.00%
GTAB	65.52%	**37.93%**	44.83%	65.52%	**37.93%**