# An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification

Ashnil Kumar, *Member, IEEE,* Jinman Kim, *Member, IEEE,*
David Lyndon, Michael Fulham, and Dagan Feng, *Fellow, IEEE*

***Abstract*—The availability of medical imaging data from clinical archives, research literature, and clinical manuals, coupled with recent advances in computer vision offer the opportunity for image-based diagnosis, teaching, and biomedical research. However, the content and semantics of an image can vary depending on its modality and as such the identification of image modality is an important preliminary step. The key challenge for automatically classifying the modality of a medical image is due to the visual characteristics of different modalities: some are visually distinct while others may have only subtle differences. This challenge is compounded by variations in the appearance of images based on the diseases depicted and a lack of sufficient training data for some modalities. In this paper, we introduce a new method for classifying medical images that uses an ensemble of different convolutional neural network (CNN) architectures. CNNs are a state-of-the-art image classification technique that learns the optimal image features for a given classification task. We hypothesise that different CNN architectures learn different levels of semantic image representation and thus an ensemble of CNNs will enable higher quality features to be extracted. Our method develops a new feature extractor by fine-tuning CNNs that have been initialised on a large dataset of natural images. The fine-tuning process leverages the generic image features from natural images that are fundamental for all images and optimises them for the variety of medical imaging modalities. These features are used to train numerous multi-class classifiers whose posterior probabilities are fused to predict the modalities of unseen images. Our experiments on the ImageCLEF 2016 medical image public dataset (30 modalities; 6776 training images and 4166 test images) show that our ensemble of fine-tuned CNNs achieves a higher accuracy than established CNNs. Our ensemble also achieves a higher accuracy than methods in the literature evaluated on the same benchmark dataset and is only overtaken by those methods that source additional training data.**

***Index Terms*—deep learning, convolutional neural network, fine-tuning, ensembles, image classification**

## I. Introduction

A diverse range of imaging data are acquired in modern hospitals for diagnosis, treatment planning, and assessing response to treatment. These large image collections coupled with other image sources (e.g., research literature and clinical manuals) provide new opportunities to use massive image data to derive computerised tools for image-based diagnosis, teaching, and biomedical research [1]. These applications are predicated on the identification, retrieval, and classification of patient data that represent similar clinical outcomes [2], e.g., images representing the same diagnosis. There is a global push to use all possible sources of images for these applications [3]–[5] but not all image sources are labelled appropriately.

In cases where appropriate labels are absent, identifying the image's modality is a primary step [6] because the content and semantics of an image can vary depending on its modality. Modality classification is inherently a multi-class problem where there are many different types of images and each type must be uniquely distinguished from all others. The core challenge arises from the fact that some modalities are visually quite distinct (e.g., anatomical and functional modalities) while others are only subtly different (e.g., different types of anatomical images). There are also variations in the appearance of images based on the individual diseases depicted. These challenges are further compounded by the lack of sufficient training data, especially for the numerous different diseased states, which hinders the application of most classification techniques. The distribution of the available labelled data is often skewed against those modalities that are more difficult for humans to interpret and label.

Attempts at modality classification have generally avoided these issues by manually sourcing and labelling images to expand the standard public benchmark datasets [7]. Classification is performed by extracting and fusing a multitude of image features from the expanded dataset (see Section I-A). This is a laborious task as the design process requires iterative labelling, tests, and calibrations (i.e., "hand-crafted" feature engineering). Moreover the approaches employed are often specific to a particular task and cannot be applied to different datasets or tasks. A data-driven approach for image feature design, as is now common in general image classification research, would be more robust to the variety of image modalities and diseases while being less susceptible to human domain-specific subjectivity.

Convolutional Neural Networks (CNNs) are a deep learning technique that implicitly perform feature extraction on image data with deeper networks generally learning more sophisticated representations of the image data [8]–[10]. Training CNNs to perform this kind of automated feature extraction typically comes with the onus of requiring large volumes of labelled training data. When such training corpora are available, CNNs are capable of achieving state-of-the-art performance in general object recognition, as evidenced by their dominance of the ImageNET benchmark [11]. A variety

A. Kumar, J. Kim, D. Lyndon, and D. Feng are with the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, The University of Sydney, Australia.

M. Fulham is with the Department of Molecular Imaging, Royal Prince Alfred Hospital, Australia and Sydney Medical School, The University of Sydney, Australia.

D. Feng is also with the Med-X Research Institute, Shanghai Jiao Tong University, China.

This work was supported in part by ARC grants.

of CNN architectures have been introduced and continue to be improved (see Section I-A). Individual architectures have different capabilities in their ability to characterise or represent image data, which is often linked to the depth of the CNN. However, CNNs may be indirectly limited when used with highly variable image datasets with limited samples (e.g., medical images): shallow CNNs may be too general and would not be able to capture the subtle differences between such images while deep CNNs may become highly sensitive to subtle differences and would not be able to capture the general similarity between such images.

In this paper, we describe a method for classifying the modality of medical images using an ensemble of different CNN architectures. Ensemble learning is a machine learning process in which better predictive performance is obtained by combining the results from multiple classification models into one high-quality classifier [12]. Our method resolves the challenges associated with using CNNs on multi-class classification problems with limited and unevenly distributed sample data by using CNNs that have been pre-trained on a large collection of natural images ($>1$ million) and fine-tuning (optimising) them using a smaller medical image dataset (thousands). The various CNNs in our ensemble allow us to extract image features at different semantic levels thereby enabling the characterisation of the varying distinct and subtle differences among modalities. Our ensemble of fine-tuned CNNs allows us to adapt the generic features learned from natural images to be more specific for different medical imaging modalities.

### A. Related Work

Outcomes from the ImageCLEF medical benchmarks [5] indicate that modality-specific information can improve the performance of image-based classification and retrieval algorithms [13]. As such, modality classification remains an important research task [14], [15].

Many prior research studies [7], [16]–[20] into modality classification have used a variety of approaches that combine a vast range of image features that were derived both globally over the whole image and locally over several different sub-patches. These works all used combinations of image features that were designed by humans to represent some characteristic of the underlying image data, e.g., textures, colours, binary patterns, and key point descriptors. The performance of these methods was implicitly tied to the quality of the features, the optimisation of which may require domain experts to hand-craft the image features. Many of these methods also used manual dataset expansion to increase the size of their training dataset, which may not be possible in real world contexts [21].

CNNs are the state-of-the-art deep learning method for image classification as demonstrated by their dominance of the ImageNet benchmark [11]. A variety of different architectures have been introduced for the classification of the 1000 categories in the ImageNet dataset [22]. The initial landmark breakthrough of Krizhevsky et al. [8] was achieved by an efficient GPU implementation of their AlexNet CNN. To reduce the risk of overfitting, Szegady et al. [23] introduced the GoogLeNet architecture, which used networks-within-networks to achieve a $12\times$ reduction in the number of parameters compared to AlexNet. Recently, He et al. [24] introduced the Deep Residual Network (ResNet) architecture, which solved the problem of degradation of training accuracy in very deep networks. While these very deep networks have high accuracy, they require several weeks to train optimally even when using state-of-the-art computing hardware.

CNNs generally require large training datasets (tens of thousands if not millions) and as such their direct application to medical imaging is difficult due to the time and labour cost involved in creating expertly labelled training datasets. Anthimopoulos et al. [25] showed that CNN architectures have higher accuracy than other methods when significant effort has been expended to acquire labels for the training data. However, when only small training datasets are available, which is the norm, CNN-based methods may overfit and struggle to learn the best image features, e.g., overfitting when only 500 images are used [26]. A recent study [27] has shown that transfer learning can be used to adapt CNNs for medical imaging. Transfer learning is the process by which a CNN is initially trained on a large well-labelled natural image dataset to learn generic image features applicable to all images and is then used to extract these generic features from smaller datasets [28]. It has been successfully utilised in various modality or disease specific studies [29]–[31]. However, the transfer learned features are more reflective of the natural image dataset and may not necessarily reflect the subtle characteristics of medical images.

A more advanced form of transfer learning, called fine-tuning, can be used to adapt a pre-trained CNN to a different dataset. Fine-tuning is the process of updating the pre-trained weights of a CNN through the use of backpropagation. An extensive study on medical imaging data has demonstrated that fine-tuning is as effective as training a CNN from scratch while being more robust to the size of training data [32]. Fine-tuning has been applied to a variety of different medical imaging modality or disease-specific classification tasks including MRI view detection [33] and ultrasound anatomy identification [34].

We have used CNNs for our preliminary work in modality classification: we have designed and trained a new CNN from scratch [35] and fine-tuned AlexNet [36]. However, both of these methods used one architecture potentially limiting their ability to extract features learned by different CNNs. Other CNN-based methods for modality classification [37], [38] have also been proposed; the best performing methods generally sourced additional images to expand the training dataset as a way to address the difficulty of learning from an unevenly distributed dataset [37].

## II. Materials

We used the medical Subfigure Classification dataset from the public ImageCLEF 2016 collection[1]. The dataset contained 6776 training images and 4166 test images across 30 different modalities. Ground truth annotations were provided for both datasets. Table VI shows the distribution of the modalities in the datasets, divided according to groups of similar modalities. The training and test sets are both skewed, with over two-thirds

[1]http://www.imageclef.org/2016/medical

of the classes having less than 100 samples. The distribution of samples was intended to reflect the availability of labelled training data. A detailed description of the datasets can be found in the ImageCLEF 2016 overview papers [15], [39].

### A. Data Augmentation

Data augmentation is the most common method used to reduce overfitting during CNN training by artificially enlarging the dataset using class-preserving perturbations of individual images [8]. The key concept is that the reproducible perturbations applied to the data do not change the semantic meaning of the image, thereby enabling the generation of new samples. Training CNNs on this larger perturbed dataset has been shown to improve robustness and generalisability to unseen data [8]. Data augmentation is a contrast to the manual sourcing of additional labelled images, which is difficult in the medical domain (Section I).

We used an established 10-fold augmentation scheme involving cropping and flipping (reflection) of each image [40]:

1) We set the dimensions of the crop to be the input dimensions of the CNN architecture.
2) We extracted 5 crops from each image: the 4 corners and the centre.
3) We generated 5 additional samples by flipping (reflecting) each crop about the x-axis.

This scheme generated 67760 augmented training crops from the 6776 training images. A random selection of 90% of the augmented set was used for CNN fine-tuning and the remaining 10% for CNN validation. All 67760 training crops were used to train our ensemble classifier (see Section III-D).

It is important to distinguish data augmentation with classical cross-validation in the context of CNNs. The aim of augmentation is to avoid overfitting the millions of CNN filter weights to a small dataset. Cross-validation is generally used to select the optimal CNN training parameters (see Section III-C) but is often not used because it is computationally expensive given that CNN training can take an extensive amount of time.

## III. Methods

### A. Overview

Figure 1 shows an overview of our ensemble method. We first fine-tuned the CNN architectures that had been pre-trained (initialised) on natural image data. Each of the fine-tuned CNNs was then used in two ways: (i) as an image feature extractor with the independent feature vectors concatenated and used to train multi-class support vector machines (SVMs), and (ii) as a classifier generating softmax probabilities. The posterior probabilities from the ensemble of SVMs and softmax classifiers were used to determine the class of the image.

We used the following different CNN architectures, each with their own different capabilities:

1) **AlexNet** [8]. This well-established CNN follows a standard neural network architecture of stacked and connected layers. It comprises eight layers that need to be trained, five convolutional layers followed by three fully-connected layers, as well as max-pooling layers. The first, second, and fifth convolutional layers are followed by overlapping max-pooling layers that make it more difficult for the network to overfit. The output of the fifth convolutional layer (after max-pooling) is fed into the stack of fully-connected layers. A rectified linear unit (ReLU) non-linearity is applied to each convolutional and fully connected layer to enable faster training.
2) **GoogLeNet** [23]. This CNN architecture introduced a new "Inception" module, a subnetwork comprising of parallel convolutional filters whose outputs are concatenated. The repetition of the Inception modules captures the optimal sparse representation of the image while simultaneously reducing dimensionality. The network comprises 22 layers that require training (or 27 if pooling layers are also considered). Experiments have shown that GoogLeNet has fewer trainable weights than AlexNet and is more accurate [23].

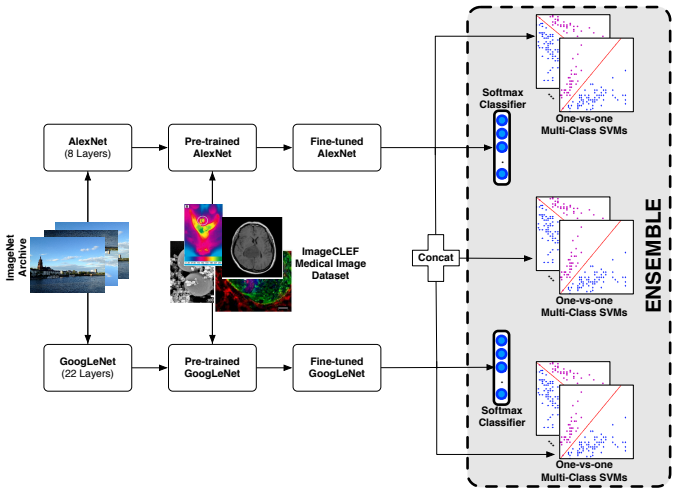


Fig. 1. Overview of our ensemble method.

We chose these architectures because they are well-established and have shown good performance when adapted to a variety of medical image classification scenarios [32], [33], [41].

We used two types of classifiers within our ensemble:

1) **softmax**. The softmax function is a generalisation of the logistic function that highlights the largest values in a vector while suppressing those that are significantly below the maximum. When applied to a $D$-dimensional feature vector, the softmax function can be used as a non-linear variant of multinomial logistic regression to generate a vector of $D$ probability values, the $d$-th element of which is the likelihood that the vector represents a member of the $d$-th class [42]. The softmax function is widely used as the classification layer of many CNN architectures [8], [23], [24].
2) **one-vs-one multi-class SVMs**. SVMs [43] are a well-established supervised binary classification technique, where the model divides labelled training data into two categories and classifies new samples into one of these. Multi-class problems are usually solved by combining multiple SVMs and as such we trained $A$-vs-$B$ (one-vs-one) SVMs for every pair of image modalities $A, B$ in

our dataset. The posterior probabilities were estimated by minimising the Kullback-Leibler divergence based on the outputs of the individual one-vs-one SVMs [44].

We implemented our method in MATLAB, using the MatConvNet library [45] for our implementation of CNN fine-tuning. For our experiments, we used the pre-trained CNNs provided with MatConvNet.

### B. CNN Fine-Tuning

The CNN architectures we used were pre-trained (initialised) on the ImageNet [11], [22] natural image dataset, which contains 1000 classes across $> 1$ million samples. We adapted the CNNs to our problem by replacing the last fully connected layer (intended for 1000 classes) with a new fully connected layer for the 30 classes in our dataset. The initial CNN filter weights derived from the natural images were then fine-tuned (optimised) through back-propagation so that they better reflected the modalities in the medical imaging dataset.

Let $\mathbf{X}$ be the training dataset of $n$ images. Fine-tuning is an iterative process that finds the filter weights $\mathbf{w}$ that minimises the CNN's empirical loss (i.e., reduces the error rate):

$$L\left(\mathbf{w}, \mathbf{X}\right) = \frac{1}{n} \sum_{i=1}^{n} l\left(f\left(\mathbf{x}_i, \mathbf{w}\right), \hat{c}_i\right) \tag{1}$$

where $\mathbf{x}_i$ is the $i$-th image of $\mathbf{X}$, $f\left(\mathbf{x}_i, \mathbf{w}\right)$ is the CNN function that predicts the class $c_i$ of $\mathbf{x}_i$ given $\mathbf{w}$, $\hat{c}_i$ is the ground-truth class of the $i$-th image, and $l\left(c_i, \hat{c}_i\right)$ is a penalty function for predicting $c_i$ instead of $\hat{c}_i$. We set $l$ to the logistic loss function.

We used mini-batch stochastic gradient descent to find the optimal $\mathbf{w}$. Let $\mathbf{B} \subset \mathbf{X}$ be a subset of $b$ images; we call $\mathbf{B}$ a mini-batch of $\mathbf{X}$ with batch size $b$. We now introduce a distinction between the terms *epoch* and *iteration*. An *epoch* is one training pass (weight update) using all the training samples in $\mathbf{X}$. In contrast, an iteration is one training pass over all the elements of $\mathbf{B}$. Each epoch, a randomized set of disjoint mini-batches were generated such that all the elements of $\mathbf{X}$ were covered. Generally speaking, one epoch consists of $\frac{n}{b}$ mini-batches of size $b$. However, when $b$ is not a factor of $n$ the last mini-batch may have less than $b$ images.

We iterated over the mini-batches of each epoch; the CNN weights were updated each iteration. The updated weights $\mathbf{w}_{t+1}$ were calculated from the gradient of the loss $L$ when applied to the mini-batch $\mathbf{B}$ using the current weights $\mathbf{w}_t$:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \left[\alpha \Delta \mathbf{w}_t - \frac{\partial L\left(\mathbf{w}_t, \mathbf{B}\right)}{\partial \mathbf{w}_t} - \lambda \mathbf{w}_t\right] \tag{2}$$

where $\Delta \mathbf{w}_t = \mathbf{w}_t - \mathbf{w}_{t-1}$ is the weight update from the previous iteration. The coefficient $\eta$ is the learning rate controlling the size of the updates to the weights. The momentum coefficient $\alpha$ diminishes fluctuations in weight changes over consecutive iterations by adding a proportion of the previous update to the current update; this has the effect of speeding up the learning process while simultaneously smoothing the weight updates. The weight decay $\lambda$ shrinks the weights to find the smallest optimal weights. In our fine-tuning setup, $\mathbf{w}_0$ are the weights of the pre-trained CNNs with $\Delta \mathbf{w}_t = 0$. Section III-C describes the selection of these parameters.

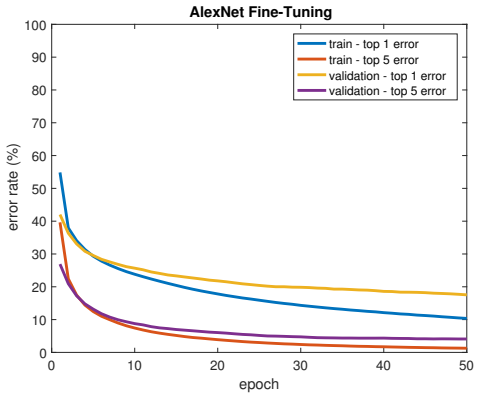


Fig. 2. Fine-tuning AlexNet over 50 epochs.

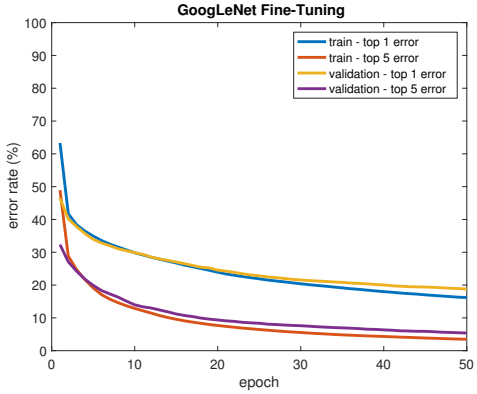


Fig. 3. Fine-tuning GoogLeNet over 50 epochs.

TABLE I
FINE-TUNING STATISTICS

| Architecture | # weights | time (s) |
|---|---|---|
| AlexNet | 56,991,134 | 14,722 |
| GoogLeNet | 6,004,302 | 39,394 |

Figures 2 and 3 show the training and validation error when fine-tuning our CNNs across the 50 epochs. For both CNNs, there is a consistent pattern of a steady drop and plateauing of the training error rates. The similarity of the training and validation curves suggests that our fine-tuned CNNs did not overfit to the training data. Table I shows the training statistics for each architecture: the number of weights that need to be fine-tuned and the total time to fine tune over 50 epochs.

### C. Fine-Tuning Parameter Selection

For the purposes of comparative evaluation we used the same parameter set for all of the CNNs in our ensemble instead of optimising the parameters for individual architectures. We used empirical methods to determine the parameters as described below. We did not use cross-validation to determine these parameters because it was computationally infeasible given the large parameter space and the time (over half a day) required to train both AlexNet and GoogLeNet given *a single* parameter combination (see Table I).

As mentioned in Section III-B we trained our CNNs for 50 epochs, given the parameter set described below. This was because the error rates for training and validation had already begun to plateau at 50 epochs (see Figures 2 and 3) with minimal improvements even when we trained for more epochs.

Each architecture required memory to store the filter weights $\mathbf{w}$ and as such the batch size $b$ was dependent on the memory capacity of the training hardware. We set $b = 256$, which was the power of 2 that maximized the memory usage of our 12GB NVIDIA Titan X GPU.

We selected a uniform learning rate $\eta = 5 \times 10^{-6}$ that allowed the fine-tuning process to effectively learn the filter weights for the different CNN architectures. We empirically determined this value by monitoring the validation error during fine-tuning using a variety of learning rates sourced from our preliminary work [26], [35], [36]; the learning rate was modified until an appropriate value was discovered. Higher learning rates often led to overfitting while lower rates led to limited change in error across epochs (i.e., slow learning).

Our momentum term $\eta\alpha\Delta\mathbf{w}_t$ controlled the fluctuation of the weights by adding a proportion of the change from the previous iteration to the current iteration (see Equation 2). As such, higher values of $\alpha$ reduced fluctuation by forcing weights to change in a similar direction to the previous iteration, leading to a smoother and faster convergence to the optimal weights. Lower values of $\alpha$ are generally used during earlier iterations when the learning process may not be globally optimal and drastic changes are more acceptable. We selected $\alpha = 0.9$ for all epochs because we were using pre-trained CNNs that had already been trained on an extensive image dataset and whose pre-trained weights would thus be appropriate (but not yet ideal) for a wide range of image data.

The weight decay term $-\eta\lambda\mathbf{w}_t$ acted as a regularisation term for the gradient descent by preventing the weights from growing too large; it was also important for avoiding overfitting. We used the default value of $\lambda = 1$.

### D. Ensemble Design

Our ensemble comprised of the following classifiers:

1) Fine-tuned AlexNet using a softmax classifier.
2) Fine-tuned GoogLeNet using a softmax classifier.
3) A one-vs-one multi-class SVM trained using features extracted from the fine-tuned AlexNet. We extracted 4096 features using the activations of the last fully connected layer of the fine-tuned network. For efficient classifier training, we reduced the dimensionality using Principle Component Analysis (PCA) [46]. Our feature vectors were the principle components that explained 90% of the variation in the data (dimensionality: 459).
4) A one-vs-one multi-class SVM trained using features extracted from the fine-tuned GoogLeNet. We extracted 1024 features using the activations of the last pooling layer of the fine-tuned network. As above, we used PCA so that the features were the components that explained 90% of the data variation (dimensionality: 108).
5) A one-vs-one multi-class SVM trained using features extracted from the fine-tuned AlexNet and GoogLeNet. The features captured from each CNN were concatenated to form a single 5120-dimensional vector and then reduced to the principle components that explained 90% of the data variation (dimensionality: 508).

The multi-class SVMs in our ensemble were trained using the PCA-reduced features extracted from all augmented variations of the training dataset. We used one-vs-one multi-class SVM classifiers to maximise the ability to distinguish two modalities that have very subtle differences (i.e., are highly similar). During classification, we first generated local crops of a given test image using the same process as data augmentation (Section II-A). We then obtained the posterior probability $P_{i,k}(m)$ that the $i$-th crop of the test image depicted a particular modality $m$ according to the $k$-th classifier in the ensemble. We determined the modality $m^\star$ of an image by fusing the posterior probabilities according to:

$$m^\star = \arg\max_m \frac{\sum_i^A \sum_k^C P_{i,k}(m)}{A \times C} \tag{3}$$

where $A = 10$ is the number of augmented variations of each test image and $C = 5$ is the number of ensemble classifiers.

## IV. Evaluation

### A. Experimental Setup

We compared our new ensemble method to a variety of well-established CNN-based methods:

- transfer learned CNNs with multi-class SVMs [28].
- fine-tuned CNNs with softmax.
- fine-tuned CNNs with multi-class SVMs.

We used the AlexNet and GoogLeNet architectures for all of these baselines. For the baseline companions, we used the standard performance measures in CNN studies [8], [23]: the correctness of the predicted label (Top 1 Accuracy) and the presence of the correct label among the 5 labels with the highest probability (Top 5 Accuracy).

We also compared the Top 1 Accuracy of our ensemble with other studies using the ImageCLEF 2016 benchmark dataset. It is important to note that several studies manually expanded the skewed training dataset with labelled images from other sources, while we used only the specified training dataset.

We also analysed the ability of our ensemble to classify individual classes within the unbalanced dataset. We measured the precision (positive predictive value or proportion of true positives), sensitivity (true positive rate), specificity (true negative rate), and F-score for each of the modalities in our dataset.

### B. Results

Table II compares the Top 1 Accuracy of our method to the CNN baselines; our method achieved a higher accuracy than all the other methods. Similarly, Table III shows the Top 5 Accuracy of our method compared to the baselines; our method achieved an accuracy that was consistent with the best accuracies among the other methods. Our method had a higher accuracy than softmax classification in both the Top 1 and Top 5 classification. In Top 1 Accuracy, our method was 1.73% more accurate than the best baseline method (fine-tuned GoogLeNet with SVM); in Top 5 Accuracy, our method

TABLE II
TOP 1 CLASSIFICATION ACCURACY (%)

| | Architecture | |
|---|---|---|
| | **AlexNet** | **GoogLeNet** |
| transfer learned + SVM | 79.21 | 78.61 |
| fine-tuned + softmax | 79.62 | 77.17 |
| fine-tuned + SVM | 79.60 | 80.75 |
| our ensemble | 82.48 | |

TABLE III
TOP 5 CLASSIFICATION ACCURACY (%)

| | Architecture | |
|---|---|---|
| | **AlexNet** | **GoogLeNet** |
| transfer learned + SVM | 96.71 | 96.33 |
| fine-tuned + softmax | 94.48 | 91.31 |
| fine-tuned + SVM | 96.47 | 96.54 |
| our ensemble | 96.59 | |

TABLE IV
COUNT OF IMAGES CORRECTLY CLASSIFIED BY A PAIR OF METHODS

| | | AlexNet[a] | | | GoogLeNet[a] | | | |
|---|---|---|---|---|---|---|---|---|
| | | **TLS** | **FTC** | **FTS** | **TLS** | **FTC** | **FTS** | **ENS** |
| **Alex** | **TLS** | | 3087 | 3113 | 3047 | 3000 | 3102 | **3188** |
| | **FTC** | 3087 | | 3118 | 3024 | 3036 | 3107 | **3231** |
| | **FTS** | 3113 | 3118 | | 3038 | 2991 | 3086 | **3211** |
| **GoogLe** | **TLS** | 3047 | 3024 | 3038 | | 2974 | 3085 | **3142** |
| | **FTC** | 3000 | 3036 | 2991 | 2974 | | 3065 | **3132** |
| | **FTS** | 3102 | 3107 | 3086 | 3085 | 3065 | | **3235** |
| | **ENS** | **3188** | **3231** | **3211** | **3142** | **3132** | **3235** | |

[a] TLS = transfer learned + SVM, FTC = fine-tuned + CNN softmax, FTS = fine-tuned + SVM, ENS = our ensemble

TABLE V
TOP 1 ACCURACY COMPARED TO OTHER METHODS

| **Method** | **Accuracy (%)** |
|---|---|
| transfer learned ResNet-152 [37]* | 85.38 |
| hand-crafted feature collection [37]* | 84.46 |
| RGB color PHOW [20]* | 84.01 |
| *our ensemble* | *82.48* |
| RGB color PHOW [20] | 81.73 |
| modified GoogLeNet (60 epochs) [37]* | 81.03 |
| fine-tuned AlexNet (100 epochs) [36] | 77.55 |
| VGG-like CNN (500 epochs) [38] | 65.31 |

* training dataset expanded with additional examples

was 0.12% lower than the best performing method (transfer learned AlexNet with SVM).

Table IV demonstrates that our ensemble had a high degree of similarity to every baseline. Each cell shows the number of images correctly classified by both of the methods noted in the row and column headers. The highest value (in bold) for every baseline was with our ensemble, indicating that the ensemble possessed the strengths of the individual methods.

Table V compares the Top 1 Accuracy of our ensemble to other image-based methods (no text data used). The methods marked with an asterix (∗) manually expanded the skewed ImageCLEF 2016 training dataset with labelled images from other sources. Our ensemble had a higher accuracy than all

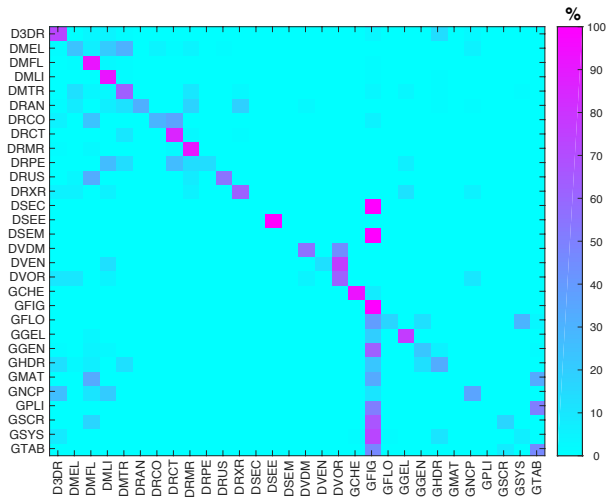


Fig. 4. Confusion matrix for our ensemble. The matrix entries have been scaled to a percentage to account for the uneven distribution of the classes.

methods that used the unexpanded dataset.

Table VI shows the precision, sensitivity, specificity, and F-score of our ensemble in the classification of the individual modalities. This is complemented by Figure 4, which is a heat map of the confusion matrix for the multi-class classification. Due to the skewed test dataset, we generated the confusion matrix by scaling the classification counts to a percentage.

### C. Discussion

Our findings show that our method achieved a higher Top 1 Accuracy than all the other baselines. This is attributed to two complementary components of our method: (i) fine-tuning to learn features that were specific to our dataset, and (ii) our use of different architectures, which had different capabilities in generalising and adapting to different data.

The transfer learned architectures extracted generic image features relevant to all images and thus could not achieve a better Top 1 Accuracy because their features were not specific for medical image modality classification. GoogLeNet, which had higher accuracy than AlexNet in the classification of natural images [23], had a lower Top 1 Accuracy when transferred to medical images. We suggest that this outcome indicates that shallower networks, such as AlexNet, learn more generalisable features that are applicable to a wider variety of images. In contrast, deeper networks, such as GoogLeNet, learn more semantically meaningful features [9] that are limited in their applicability when transferred to a different domain. Since transfer learned AlexNet features are less semantically optimised for natural images (in comparison to the deeper GoogLeNet), they are more generalisable and adaptable when transferred to the medical imaging domain. The transfer learning outcomes demonstrate the characteristics and strengths of the different architectures.

Using CNN softmax classification (the standard for most natural image classification tasks) after fine-tuning revealed some important outcomes. For AlexNet there was improvement in the classification accuracy (0.41%) in comparison to transfer learned AlexNet; in contrast, the accuracy of

TABLE VI
PER CLASS CLASSIFICATION RESULTS (%)

| Group | Class | # Samples | | Results | | | |
|---|---|---|---|---|---|---|---|
| | | Train | Test | Precision | Sensitivity | Specificity | F-Score |
| | 3D reconstructions (D3DR) | 201 | 96 | 69.31 | 72.92 | 99.24 | 71.07 |
| microscopy | electron microscopy (DMEL) | 208 | 88 | 39.62 | 23.86 | 99.22 | 29.79 |
| | fluorescence microscopy (DMFL) | 906 | 284 | 73.45 | 91.55 | 97.58 | 81.50 |
| | light microscopy (DMLI) | 696 | 405 | 87.94 | 91.85 | 98.64 | 89.86 |
| | transmission microscopy (DMTR) | 300 | 96 | 48.39 | 62.50 | 98.43 | 54.55 |
| radiology | angiography (DRAN) | 17 | 76 | 92.31 | 31.58 | 99.95 | 47.06 |
| | combined modalities (DRCO) | 33 | 17 | 41.67 | 29.41 | 99.83 | 34.48 |
| | computerised tomography (DRCT) | 61 | 71 | 80.26 | 85.92 | 99.63 | 82.99 |
| | magnetic resonance (DRMR) | 139 | 144 | 75.29 | 90.97 | 98.93 | 82.39 |
| | positron emission tomography (DRPE) | 14 | 15 | 100 | 13.33 | 100 | 23.53 |
| | ultrasound (DRUS) | 26 | 129 | 98.59 | 54.26 | 99.98 | 70.00 |
| | x-ray, 2D radiography (DRXR) | 51 | 18 | 34.38 | 61.11 | 99.49 | 44.00 |
| signals | electrocardiography (DSEC) | 10 | 8 | 0 | 0 | 100 | 0 |
| | electroencephalography (DSEE) | 8 | 3 | 100 | 100 | 100 | 100 |
| | electromyography (DSEM) | 5 | 6 | 0 | 0 | 100 | 0 |
| photos | dermatology, skin (DVDM) | 29 | 9 | 62.50 | 55.56 | 99.93 | 58.82 |
| | endoscopy (DVEN) | 16 | 8 | 100 | 12.50 | 100 | 22.22 |
| | other images (DVOR) | 55 | 21 | 52.00 | 61.90 | 99.71 | 56.52 |
| generic biomedical illustrations | chemical structure (GCHE) | 61 | 14 | 92.86 | 92.86 | 99.98 | 92.86 |
| | statistics, figures, graphs, charts (GFIG) | 2954 | 2085 | 88.80 | 99.23 | 87.46 | 93.73 |
| | flowcharts (GFLO) | 20 | 31 | 71.43 | 16.13 | 99.95 | 26.32 |
| | chromatography, gel (GGEL) | 344 | 224 | 95.03 | 76.79 | 99.77 | 84.94 |
| | gene sequence (GGEN) | 179 | 150 | 71.74 | 22.00 | 99.68 | 33.67 |
| | hand-drawn sketches (GHDR) | 136 | 49 | 29.09 | 32.65 | 99.05 | 30.77 |
| | mathematics, formula (GMAT) | 15 | 3 | 0 | 0 | 100 | 0 |
| | non-clinical photos (GNCP) | 88 | 20 | 36.84 | 35.00 | 99.71 | 35.90 |
| | program listing (GPLI) | 1 | 2 | 0 | 0 | 100 | 0 |
| | screenshots (GSCR) | 33 | 6 | 50.00 | 16.67 | 99.98 | 25.00 |
| | system overviews (GSYS) | 91 | 75 | 33.33 | 6.67 | 99.76 | 11.11 |
| | tables and forms (GTAB) | 79 | 13 | 50.00 | 46.15 | 99.86 | 48.00 |

GoogLeNet dropped by more than 1%. The reason for this is that the softmax classifier does not conduct a one-vs-one comparison when classifying an image and as such it may be possible for modalities with subtle differences to be misclassified; a one-vs-one multi-class SVM is capable of distinguishing images with these subtle differences.

Fine-tuning the networks and then using SVMs improved the Top 1 Accuracy for both the networks. It is worth noting that the accuracy of GoogLeNet improved more than that of AlexNet. This suggests that AlexNet is less prone to be fine-tuned even when using the same fine-tuning parameters and data. Additionally it suggests that the features learned by the deeper GoogLeNet were more semantically relevant for the medical images, thereby introducing a stronger discriminative capability. These findings are important because they indicate that deeper networks are more likely than shallower networks to learn relevant features when fine-tuned on a smaller dataset.

Our method achieved a competitive Top 5 Accuracy; it was higher than all of the other methods except for transfer learned AlexNet, where it was 0.12% lower (approximately 5 images whose true modality was not reflected in the Top 5). This is an interesting outcome given that the overall Top 1 Accuracy increased compared to the other methods. Our explanation is that the fine-tuning of our ensemble may have indirectly weakened the ability to extract image features relevant to this relatively small subset of 5 images because the fine-tuning process tried to extract image features that maximised the overall classification accuracy (see Section III-B). In contrast, transfer learning extracts generic image features that are applicable to all images (see Section I-A) and as such the transfer learned AlexNet was able to predict the correct modality in the Top 5. However, because the generic image features were not optimised for the medical image dataset, the transfer learned AlexNet had an overall lower Top 1 Accuracy.

The confusion matrix (Figure 4) and the numerical outcomes (Table VI) indicate that many of the modalities were correctly identified: 19 of the 30 modalities had precision $\geq 50\%$, 14 modalities with sensitivity $\geq 50\%$, and 13 modalities with F-score $\geq 50\%$. The confusion matrix indicates that misclassifications generally occurred within a group (e.g., within the radiology group, DRCO was misclassified as DRCT) but there were situations when misclassifications occurred outside the group, most commonly with GFIG. This is attributed to GFIG being the largest and most varied class in the dataset.

The classes that received 0 precision and F-score were those classes with: (i) less than 20 training samples, and (ii) subtle differences shared with other classes with larger training sets, e.g., DSEC signals are visually similar to line graphs in GFIG. Other classes with less than 20 samples (e.g., DSEE) had higher F-scores due to visually distant characteristics, e.g., colour and position of signals. Similar patterns also occur for other classes with few samples and very subtle differences. For example, the DRCO class (33 training images) contained multi-modality images and these were sometimes misclassified as one of the constituent modalities, e.g. DRCT. This indicates that it may be important to increase the number of training samples for classes with subtle differences, as done in other work [20], [37]. Alternatively, the training loss function (see Equation 1) could be scaled according to the number of samples in each class but this requires a separate study as it may in some cases deteriorate classification performance [38].

Our ensemble achieved a higher Top 1 Accuracy than other methods using hand-crafted features [20] as well as those using AlexNet and GoogLeNet [36]–[38] (see Table V). In particular, our ensemble of CNNs fine-tuned for only 50 epochs had higher accuracy than methods that used individual CNNs that were fine-tuned for a larger number of epochs (60-500). This finding indicates that our ensemble was able to create an accurate classifier from constituent CNNs that were not fine-tuned as extensively as in other works. Table V shows that two variants of existing methods [20], [37] achieved similar accuracy (over 81%) compared to our method (82.48%) while being methodologically simpler. However, it is important to note that the RGB PHOW method [20] used hand-crafted features, which are not robust to changes in the dataset and are susceptible to human domain-specific subjectivity, while the modified GoogLeNet [37] used manual dataset expansion, which requires extensive human effort and is not replicable across applications. We suggest that the ability of our method to learn from the underlying data makes it more adaptable to dataset changes. Table V also lists methods that reported higher Top 1 Accuracy. These methods expanded their training dataset with other image data, including adjusting the distribution of the classes with few samples [20], [37], and as such it is difficult to conduct a direct comparison with our ensemble because we cannot ascertain whether the improved accuracy was due to their different methods or due to their use of an expanded training dataset. Valavanis et al. [20] showed that expanding the dataset improved the accuracy of their method. We expect that our ensemble would also improve with additional training data. In particular, we expect significant improvements when our ensemble is fine-tuned with additional data for the worst performing classes (DSEC, DSEM etc.) as these are generally have higher rates of misclassifications.

Our ensemble has the ability to discriminate between distinct and subtle differences between image modalities because it merges the generalisability of the shallower AlexNet with the semantic relevance of the fine-tuned GoogLeNet. This finding is supported by Table IV, which shows that the modality classifications performed by our ensemble had high similarity to every baseline method. This result, in combination with the highest Top 1 Accuracy, indicates that the ensemble possessed the strengths of the individual methods and was thus capable of classifying images that would have been misclassified by individual methods. Furthermore, fine-tuning of the CNNs in our ensemble enables us to extract image features that are more relevant to the dataset being classified. The ability to identify the image features that are most relevant for the classification suggests that our method can be readily adapted to a variety of multi-modality and multi-disease datasets.

In order to demonstrate the influence of the ensemble, we did not perform any parameter optimisations and used the same fine-tuning parameters for both CNNs (see Section III-C). Optimising the parameters for each CNN separately via cross-validation would give improved results. Our method can also be extended through the integration of different CNN architectures with new capabilities, such as the ResNet [24], which can addresses training accuracy degradation in very deep networks. Koitka et al. [37] showed that transfer learned ResNet has higher accuracy than GoogLeNet using an expanded dataset. We believe that including ResNet as part of our ensemble would extend its capabilities. This is left to future work because training ResNet requires several weeks even when performed on multiple devices in parallel.

## V. Conclusions

In this paper, we introduced a new ensemble method for the classification of the modality of medical images. Our ensemble used multiple fine-tuned CNNs as optimised feature extractors that were able to learn image features that captured the diverse information present in medical images of different modalities. The ensemble fused the fine-tuned CNN models to derive a more powerful image classification scheme than the individual CNNs. Our experimental results showed that our ensemble was able to correctly classify the majority of images in a public benchmark dataset and achieved higher classification accuracy than other CNN baselines as well as other methods using the same benchmark training dataset.

Our experiments showed that our ensemble method was able to distinguish between image modalities with subtle differences under the constraint that there were sufficient training samples to learn the differences between the modalities. In the future, we will investigate adaptations to our fine-tuning scheme (e.g., changes to the loss function) that can potentially reduce this constraint so that our ensemble can be used without manual dataset expansion.

## Acknowledgment

The authors acknowledge NVIDIA Corporation for their donation of the Titan X GPU used for this research.

## References

[1] A. Kumar, J. Kim, W. Cai, and D. Feng, "Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data," *J Digit Imaging*, vol. 26, no. 6, pp. 1025–1039, 2013.

[2] S. Sedghi, M. Sanderson, and P. Clough, "How do health care professionals select medical images they need?" *Aslib Proc*, vol. 64, no. 4, pp. 437 – 456, 2012.

[3] R. J. Stanley, S. De, D. Demner-Fushman, S. Antani, and G. R. Thoma, "An image feature-based approach to automatically find images for application to clinical decision support," *Comput Med Imag Grap*, vol. 35, no. 5, pp. 365 – 372, 2011.

[4] S. P. Rowe, A. Siddiqui, and D. Bonekamp, "The key image and case log application: New radiology software for teaching file creation and case logging that incorporates elements of a social network," *Acad Radiol*, vol. 21, no. 7, pp. 916 – 930, 2014.

[5] J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, and H. Müller, "Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at imageclef 2004–2013," *Comput Med Imag Graphics*, vol. 39, pp. 55–61, 2014.

[6] D. Keysers, J. Dahmen, H. Ney, B. B. Wein, and T. M. Lehmann, "Statistical framework for model-based image retrieval in medical applications," *J Electron Imaging*, vol. 12, no. 1, pp. 59–68, 2003.

[7] M. Abedini, N. C. F. Codella, J. H. Connell, R. Garnavi, M. Merler, S. Pankanti, J. R. Smith, and T. Syeda-Mahmood, "A generalized framework for medical image classification and recognition," *IBM J Res Dev*, vol. 59, no. 2/3, pp. 1:1–1:18, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv Neur In*, 2012, pp. 1097–1105.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE T Pattern Anal*, vol. 35, no. 8, pp. 1798–1828, 2013.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int J Comput Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE T Biomed Eng*, vol. 59, no. 9, pp. 2538–2548, 2012.

[13] M. M. Rahman, D. You, M. S. Simpson, S. K. Antani, D. Demner-Fushman, and G. R. Thoma, "Multimodal biomedical image retrieval using hierarchical classification and modality fusion," *Int J Multimedia Inform Ret*, vol. 2, no. 3, pp. 159–173, 2013.

[14] M. Villegas, H. Müller, A. Gilbert, L. Piras, J. Wang, K. Mikolajczyk, A. G. S. de Herrera, S. Bromuri, M. A. Amin, M. K. Mohammed, B. Acar, S. Uskudarli, N. B. Marvasti, J. F. Aldana, and M. del Mar Roldán García, "General Overview of ImageCLEF at the CLEF 2015 Labs," in *LNCS*, 2015, vol. 9283, pp. 444–461.

[15] A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller, "Overview of the ImageCLEF 2016 Medical Task," in *CLEF2016 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1609, 2016, pp. 219–232.

[16] O. Pelka and C. M. Friedrich, "FHDO biomedical computer science group at medical classication task of ImageCLEF 2015," in *CLEF 2015 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1391, 2015.

[17] I. Kitanovski, I. Dimitrovski, and S. Loskovska, "FCSE at medical tasks of ImageCLEF 2013," in *CLEF 2013 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1179, 2013.

[18] M. Abedini, L. Cao, N. Codella, J. H. Connell, R. Garnavi, A. Geva, M. Merler, Q.-B. Nguyen, S. U. Pankanti, J. R. Smith, X. Sun, and A. Tzadok, "IBM research at ImageCLEF 2013 medical tasks," in *CLEF 2013 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1179, 2013.

[19] I. Dimitrovski, D. Kocev, I. Kitanovski, S. Loskovska, and S. Džeroski, "Improved medical image modality classification using a combination of visual and textual features," *Comput Med Imag Grap*, vol. 39, pp. 14 – 26, 2015.

[20] L. Valavanis, S. Stathopoulos, and T. Kalamboukis, "IPL at CLEF 2016 Medical Task," in *CLEF 2016 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1609, 2016, pp. 413–420.

[21] G. Wang, D. Forsyth, and D. Hoiem, "Comparative object similarity for improved recognition with few or no examples," in *IEEE Proc CVPR*, 2010, pp. 3525–3532.

[22] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Proc CVPR*, 2009, pp. 248–255.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Proc CVPR*, 2015, pp. 1–9.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[25] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE T Med Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[26] D. Lyndon, A. Kumar, J. Kim, P. H. W. Leong, and D. Feng, "Convolutional neural networks for medical clustering," in *CLEF 2015 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1391, 2015.

[27] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE T Med Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Int Conf Machine Learn*, 2014, pp. 647–655.

[29] C. K. Shie, C. H. Chuang, C. N. Chou, M. H. Wu, and E. Y. Chang, "Transfer representation learning for medical image analysis," in *IEEE Conf Eng Med Biol*, 2015, pp. 711–714.

[30] H. T. H. Phan, A. Kumar, J. Kim, and D. Feng, "Transfer learning of a convolutional neural network for HEp-2 cell image classification," in *IEEE Int S Biomed Imaging*, 2016, pp. 1208–1211.

[31] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J Biomed Health Informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.

[32] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE T Med Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[33] J. Margeta, A. Criminisi, R. C. Lozoya, D. Lee, and N. Ayache, "Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition," *Comput Method Biomec*, 2016. doi: 10.1080/21681163.2015.1061448.

[34] A. Kumar, P. Sridar, A. Quinton, R. K. Kumar, D. Feng, R. Nanan, and J. Kim, "Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks," in *IEEE Int S Biomed Imaging*, 2016, pp. 791–794.

[35] D. Lyndon, A. Kumar, J. Kim, P. H. W. Leong, and D. Feng, "Convolutional neural networks for medical classification," in *CLEF 2015 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1391, 2015.

[36] A. Kumar, J. Kim, D. Lyndon, and D. Feng, "Subfigure and multi-label classification using a fine-tuned convolutional neural network," in *CLEF 2016 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1609, 2016, pp. 318–321.

[37] S. Koitka and C. M. Friedrich, "Traditional Feature Engineering and Deep Learning Approaches at Medical Classification Task of ImageCLEF 2016 ," in *CLEF 2016 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1609, 2016, pp. 304–317.

[38] D. Semedo and J. Magalhães, "NovaSearch at ImageCLEFmed 2016 Subfigure Classification Task," in *CLEF 2016 Working Notes*, ser. CEUR Workshop Proceedings, vol. 1609, 2016, pp. 386–398.

[39] M. Villegas, H. Müller, A. García Seco de Herrera, R. Schaer, S. Bromuri, A. Gilbert, L. Piras, J. Wang, F. Yan, A. Ramisa, E. Dellandrea, R. Gaizauskas, K. Mikolajczyk, J. Puigcerver, A. H. Toselli, J.-A. Sánchez, and E. Vidal, "General Overview of ImageCLEF at the CLEF 2016 Labs," in *LNCS*, 2016, vol. 9822, pp. 267–285.

[40] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[41] S. Choi, "X-ray Image Body Part Clustering using Deep Convolutional Neural Network: SNUMedinfo at ImageCLEF 2015 Medical Clustering Task," in *CLEF 2015 Working Notes*, 2015.

[42] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[43] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[44] B. Zadrozny, "Reducing multiclass to binary by coupling probability estimates," in *Adv Neur In*, 2001, pp. 1041–1048.

[45] A. Vedaldi and K. Lenc, "MatConvNet – Convolutional Neural Networks for MATLAB," in *Proc ACM Int Conf Multimedia*, 2015, pp. 689–692.

[46] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.