Biomedical image classification based on a feature concatenation and ensemble of deep CNNs

Long D. Nguyen, Ruihan Gao, Dongyun Lin, Zhiping Lin[1]

School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore

Abstract

Deep learning and more specifically Convolutional Neural Network (CNN) is a cutting edge technique which has been applied to many fields including biomedical image classification. To further improve the classification performance for biomedical images, in this paper, a feature concatenation method and a feature concatenation and ensemble method are proposed to combine several CNNs with different depths and structures. Three datasets, namely 2D Hela dataset, PAP smear dataset, and Hep-2 cell image dataset, are used as benchmarks for testing the proposed methods. It is shown from experiments that the feature concatenation and ensemble method outperforms each individual CNN, and the feature concatenation method, as well as several state-of-the-art methods in terms of classification accuracy.

Keywords

Deep convolutional neural network; transfer learning; ensemble learning; feature concatenation; biomedical image classification.

1. Introduction

Biomedical images are widely used in applications such as diagnostics, healthcare, and pharmaceutical testing. By revealing areas and objects beyond the resolution range of normal naked eyes, biomedical imaging like microscopy can provide great details of structures of the finest objects. Thanks to outstanding color reproduction and a high dynamic range of detection, biomedical images can also unfold the interior of a body and micro-structures of organizations with high fidelity. For example, pap smear images are used in pap tests to differentiate diseased tissues from normal ones and facilitate early detection of cancer (Ashtarian et al., 2017) (Jantzen, et al., 2005). Hela is a durable and prolific line of cells that has been used for various experimental observations and clinical diagnosis, ranging from the development of the polio vaccine to the study of the AIDS virus (Boland & Murphy, 2001).

Despite the great benefits that biomedical images provide for biomedical research and clinical practice, conventionally only professional researchers and specialized doctors can analyze these images. Even though experts are well trained to identify characteristic patterns in the images, such as abnormal shape and color, visual inspection is prone to challenges from both subjective and objective perspectives. For example, experts, as human beings, are inexorably susceptible to fatigue, emotional fluctuation, and other stochastic subjective biases; on the other hand, the considerable variability in biomedical images

[1] Corresponding author. E-mail address: EZPLIN@ntu.edu.sg (Z. Lin).

and possible clumping and occlusion of cells in images make the examination even harder. These challenges not only make exigent demands on the time and energy of researchers and doctors, but also possibly impair the accuracy and efficiency of image analysis and classification.

In view of the above-mentioned challenges, a growing number of machine learning approaches have been proposed to address biomedical applications, such as noise suppression (Jeon, 2017), electromyogram analysis (Ambikapathy & Krishnamurthy, 2018), and time-series clinical data interpretation (Duneja, et al., 2018). Deep learning is a new and promising approach that moves one step forward by featuring its scale and hierarchical feature learning capability. Inspired by brain structure and functions, various deep learning models have been developed, most noticeably the Convolutional Neural Network (CNN). CNN obtains the "convolved feature" by moving a sampling window along the rows and columns of an input image and compute the dot product of the sliding "window" and the corresponding pixels in the image. Compared to traditional feature extraction methods, the weights shared among the convolutional layers and the properly chosen pooling after convolving equip CNN with many advantages such as translation invariance. Some pioneering and current works applying CNN to biomedical image analysis are reviewed in (Litjens, et al., 2017). For example, (Ciresan D. et al., 2012) uses a deep neural network with max-pooling layers as a pixel classifier and segments brain membranes depicted in a stack of electron microscopy (EM) images. Ref. (Ciresan D. et al., 2013) uses a deep neural network to detect mitosis to help analyze breast cancer histology images. Ref. (Kamnitsas, et al., 2017) proposes a dual pathway, three-dimensional CNN with eleven layers in depth to implement brain lesion segmentation, and many more.

However, the CNN model usually consists of many layers and needs to be trained with millions of data samples. To facilitate wider applications, several open source pre-trained models have been trained on large datasets so that they can be directly used as a feature extractor to solve real-life problems with only limited training data. This technique of applying pre-trained networks to small datasets is called transfer learning. Ref. (Esteva, et al., 2017) uses a GoogleNet Inception v3 CNN architecture to differentiate benign and malignant melanomas and to diagnose potential skin cancer. An ensemble of AlexNet and GoogleNet has been made and used multiple classifiers including support vector machine (SVM) with Principle Component Analysis (PCA) and SoftMax (Kumar et al., 2017). This work is one of the pioneering works presenting comprehensive research findings in applying CNNs to biomedical image classification. While the techniques and results presented are promising, we feel that further improvement is necessary. This could be seen from the fact that although using deep learning techniques and fine-tuning, the method in (Kumar, et al., 2017) did not outperform some of the traditional classification methods for the datasets used in their paper. As fine-tuning of CNNs takes a lot of time and human effort and may result in overfitting when the dataset for training is small, it is desirable to develop biomedical image classification methods which avoid fine-tuning.

Built upon the work and yet to alleviate the above-mentioned limitation, we propose in this paper an improved method for biomedical image classification based on a feature concatenation and ensemble of deep CNNs. The major contributions of this work can be summarized as: (i) By combining several recently developed CNNs of various depths and structures, the proposed model takes advantage of multiple sets of convolved features and extracts their complementary information. (ii) By replacing a single delicately fine-tuned model with an ensemble of networks and by adding one hidden layer before the last soft-max layer, we avoid the fine-tuning procedure. (iii) By applying transfer learning, multiple pre-trained networks are used for small medical image datasets. A preliminary version of this paper was recently presented in a conference (Nguyen et al., 2018).

The remaining part of this paper is structured as follows. Section 2 presents our proposed methodology, key terms involved and the implementation of our model based on a feature concatenation and ensemble of deep CNNs. Section 3 illustrates our experiment setting, results and a detailed analysis. Section 4 concludes the paper.

2. Methodology

In this paper, we propose a new method for biomedical image classification based on a feature concatenation and ensemble of deep CNNs and using transfer learning. Exploiting three most recent CNN models, we propose to first concatenate the last layer of features obtained from these models as a new model. Then an ensemble technique is applied to these four models as illustrated in Fig.1 and the details of the proposed method are discussed in the following.

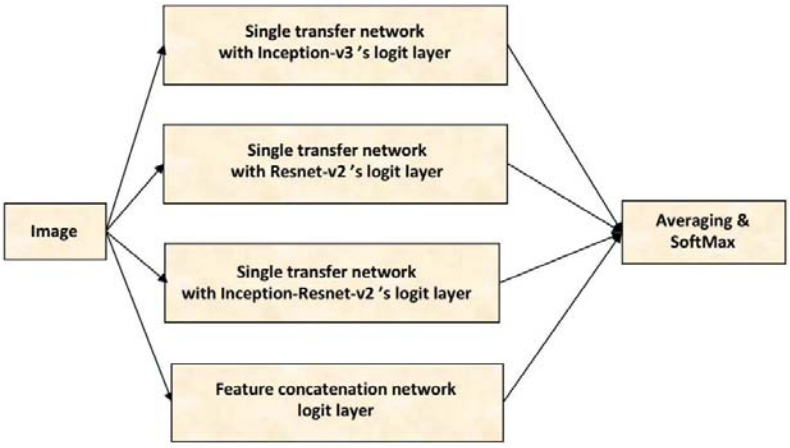


Figure 1. The structure of the proposed method

2.1 Transfer learning

Transfer learning is a machine learning technique which exploits rich and complex feature representations for smaller datasets with limited labeled samples. In transfer learning, a deep neural network is trained with a large set of labeled samples from related tasks and then used as a feature extractor for a small dataset. Given the enormous resources provided for training, the transfer learning model is expected to extract features general enough to be applied to similar but different datasets. This pre-training step with base network facilitates the subsequent learning procedure and allows more rapid progress when modelling the second and other ensuing tasks. The base model can be utilized as a feature extractor and only the last few layers of the pre-trained network need to be modified into customized layers to fit each specific task. Therefore, only the last customized layers need to be trained and the number of parameters to be adjusted is considerably reduced. These layers act as the final classifier for the dataset and small variations of the layer setting offer slightly different results. By experiments, it is observed that adding one hidden layer before the last soft-max layer provides the best result for the small biomedical image datasets we have at hand. Fig. 2 describes the structure of the transfer learning model.

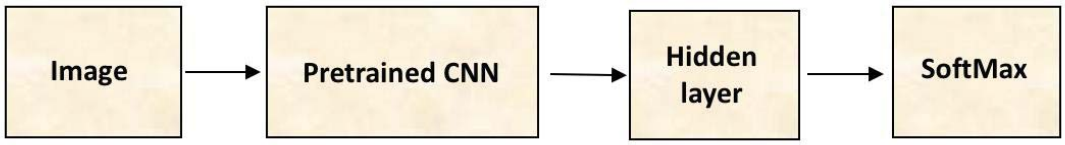


Figure 2. Transfer learning model

Thanks to ImageNet which is a large manually annotated image database, a benchmark is provided for researchers to evaluate their methods and algorithms. Many teams participate in the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and this leads to the development of various CNNs. In this paper, we choose three most recent pre-trained models trained on the ImageNet dataset, namely Inception-v3 net (Szegedy, et al., 2016), ResNet152 net (He, et al., 2016), and Inception-ResNet-v2 net (Szegedy, et al., 2017). These three models are chosen since they achieved outstanding performance in ILSVRC and represent the state-of-the-art deep learning techniques such as auxiliary classifier (Szegedy, et al., 2016), identity mapping (He, et al., 2016), label smoothing, and Rectified Linear Units (ReLU) (Szegedy, et al., 2017). In addition to these three CNNs, we propose a CNN method by concatenating the feature vectors of them. The structures of these three networks and the proposed feature concatenation are briefly described as followings:

a) Inception-v3 net (Szegedy et al., 2016)

Inception-v3 is an extended work of inception-v2 that achieves high efficiency in performing image recognition tasks by factorizing 5×5 convolution into two smaller 3×3 convolutions to speed up

computation and by expanding the filter banks in width to remove the representational bottleneck. By adding a regularizing component to the loss function, the novel Inception-v3 net attains label smoothing and to a large extent, prevents overfitting. Moreover, Inception-v3 further factorizes 7×7 convolution and concatenates multiple different layers with batch normalization technique, rendering even higher efficiency and less computational complexity. The detailed structure is shown in Fig. 3.

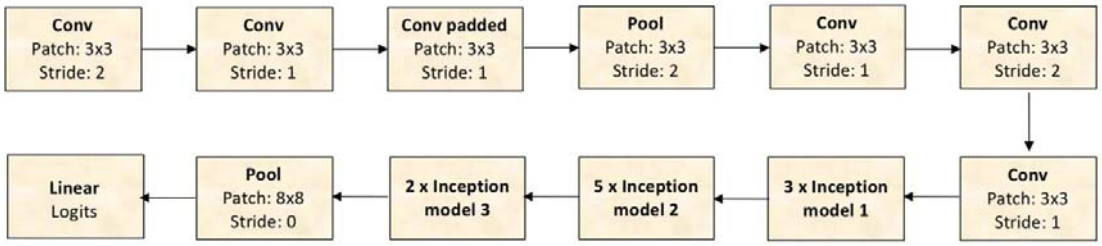


Figure 3. Inception-v3 net

b) ResNet152 net (He et al., 2016)

Inspired by Vector of Locally Aggregated Descriptors (VLAD) and residual vectors, Residual networks (Resnet) is constructed by a family of multiple deep neural networks with shortcut connections between plain networks, which implements residual learning and identity mapping. By introducing deep residual networks and recasting the traditional mapping into residual mapping, residual learning addresses the degradation problem, which is the decrease in training accuracy after reaching a saturation point. ResNet152, in particular, consists of 152 layers and is one of the deepest networks ever presented on ImageNet. It is constructed in such a delicate feedforward network with shortcut connection which skips one or more layers that it yields better classification accuracy without increasing the complexity and computational demand of the whole model. We select ResNet152 since it is one of the best performers of Resnet family members and the details of its structure are shown in Fig. 4.

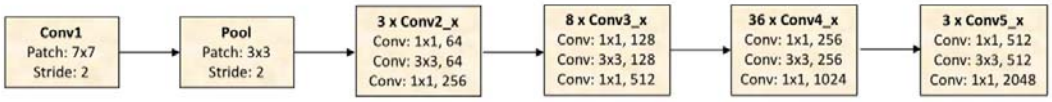


Figure 4. ResNet152

c) Inception-ResNet-v2 net (Szegedy et al., 2017)

Inception-ResNet-v2, as its name suggests, is a combination of inception net and residual net. It only performs batch-normalization on the top of traditional layers but not on the summations. This simplification reduces the overall memory footprint that is consumed and increases the total number of possible inception blocks. Furthermore, compared to previous residual variants that tend to be unstable after the number of filters exceeds 1000, Inception-ResNet-v2 manages to stabilize the training by

scaling down the residuals before adding them to previous activation layers. The details of the structure of Inception-ResNet-v2 is shown in Fig. 5.

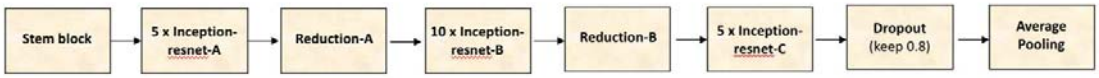


Figure 5. Inception-ResNet-v2

d) Concatenated CNN model

After applying the aforementioned neural networks and feeding with biomedical images as the input, a set of convolved features can be extracted in the form of feature map from each individual neural network. We further combine them by concatenating one feature vector after another to obtain a concatenated feature map. The fourth feature map does not require any separate network, but just compiles numerical information together and forms a new vector map. Fig. 6 shows the details of the formation of the fourth feature map.

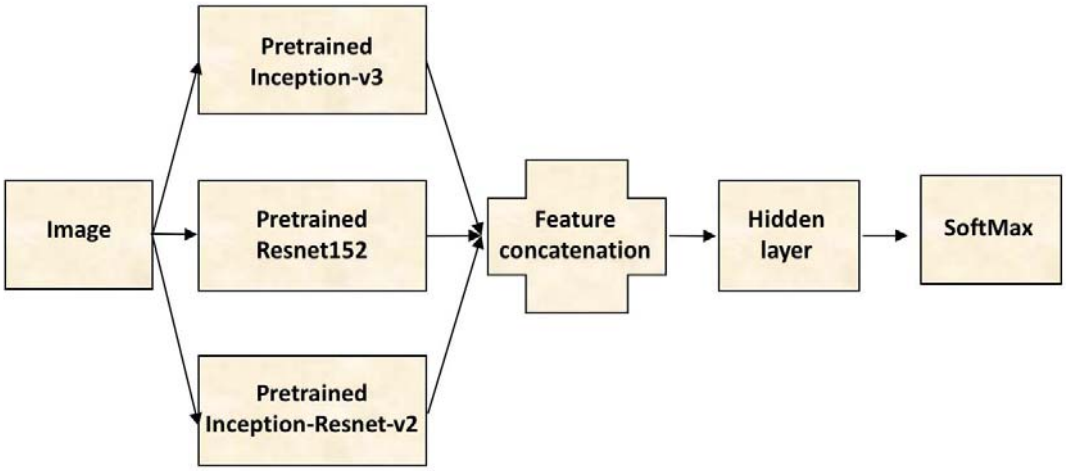


Figure 6. Concatenated CNN model

2.2 Multi-feature-extractors model

Based on the transfer learning theory, we choose three pre-trained networks, namely Inception-v3, ResNet152 and Inception-ResNet-v2 and three sets of feature maps can be extracted before the hidden layer. Note that for several research works have suggested that using various CNN features tends to improve the overall performance. In (Zheng et al., 2016), the network is designed to collect both low-level and high-level features by combining information obtained from all CNN layers. Similarly, (Kawahara & Hamrneh, 2016) makes use of images of different levels of resolution to acquire several sets of features for training the network and improves the performance on skin-lesion detection. In this paper, in addition to three feature maps obtained from pre-trained models, we add another feature descriptor that concatenates three sets of features into a longer feature vector.

2.3 Ensemble learning and the proposed overall structure

Dated back to 1990s, ensemble learning was designed to boost weaker learners to stronger ones. The method involves a number of learners called base learners and combine their results in a variety of ways, e.g. boosting, bagging and stacking (Dietterich, 1997). It has been shown that the generalization performance of ensemble learning model is significantly enhanced. Since training data can only provide limited information and it is generally impossible to find exactly one learner which always produces the best results, ensemble learning makes sense possibly because the ensemble model can always choose the optimal model in different situations and provide a better approximation of the true target function in the hypothesis space. In this paper, we ensemble the four feature maps elucidated above by taking the average of labels at the hidden layer before finally forwarding the result to the SoftMax layer for the ultimate classification. Compared to previous pioneering networks, such as AlexNet that achieves the state-of-the-art performance by averaging seven CNN models with the same structure and ResNet that champions in ILSVRC2015 by averaging six models with different depth, our new ensembled model aims to maximize the complementary effect and advantages of different networks. This combinational assembly works well and is justifiable based on Krogh and Vedelsby's theory that the base learners of the ensemble model should be as accurate and as diverse as possible (Krogh & Vedelsby, 1995).

To be specific, the proposed method combines the transfer learning models with different neural networks by an ensemble of the Inception-v3 net, ResNet152, Inception-ResNet-v2 net, and the concatenated CNN model. It is suggested in (Ju et al., 2018) that averaging multiple networks could produce a smaller variance, which is of great importance especially when the networks involved are uncorrelated to each other. The unweighted average approach is adopted, and details of implementation are shown in the following.

Consider a classification task of N classes, with M base learners. Let z_{ji} be the value of the j^{th} base learner (j = 1,2, …,M) at the i^{th}node of the last layer (i = 1,2, …,N). Since different network structures have different scaling mechanisms, values should be normalized first before performing any further computation so that the networks can be unbiasedly combined together. Let v_{ji} be the normalized value of z_{ji}:

$$v_{ji} = \frac{z_{ji} - E(z_{ji})}{\sigma(z_{ji})}$$

where $E(z_{ji})$ and $\sigma(z_{ji})$ are the expectation and standard deviation computed over the training dataset. The averaged value of all models (or the combined value) for the i^{th}node is:

$$v_i = \frac{1}{M}\sum_{j=1}^{M} v_{ji}$$

in which M is the number of the classifiers used.

Applying the SoftMax function to the final layer, the output of the network at the i^{th}node is defined as:

$$p_i = softmax(v_i) = \frac{e^{v_i}}{\sum_{k=1}^{N} e^{v_k}}$$

In our experiments, we average the logit layers given by the 4 networks (3 single transfer networks and the feature concatenation network). We then apply SoftMax to the averaged logits result to form the final prediction result. It is suggested in (Kumar et al., 2017) that fine-tuning can help transfer learning fit better to one specific task performed. However, since the dataset used in this experiment is relatively small, fine-tuning is not applied here to avoid potential overfitting. Instead, one more hidden layer is added before the final classification layer of the proposed ensemble model. This adjustment aims to extend the learning capability of the proposed network and to adapt the generic features extracted from transfer learning to the specific tasks performed without fine-tuning.

3. Experiments

Three experiments were conducted to perform the classification task on three biomedical image datasets to be discussed in this section. In the first experiment, the effectiveness of the proposed feature concatenation and ensemble method is evaluated through various combinations of networks; the performance of the proposed feature concatenation method and that of the feature concatenation and ensemble method are compared to the performance of the three individual CNNs, respectively. In the second experiment, the proposed method is compared with two recent machine learning methods using hand-craft features: (i) the spatial adjacent histogram based on adapted local binary patterns (SAHLBP) (Liu, et al., 2016) and (ii) a reject option based cascade structure of a support vector machine (SVM) with subspace analysis (Lin et al., 2018). These two methods are chosen for comparison because they are the best available published methods for the three benchmark biomedical image datasets so far. In the third experiment, the proposed method is compared with an ensemble CNN method presented in (Kumar, et al., 2017), which applies finetuning to AlexNet and GoogleNet, feeds the extracted features to Support Vector Machine (SVM) and SoftMax classifiers, and finally performs feature concatenation and ensemble five classifiers to obtain the classification results.

3.1 Datasets

Three public biomedical image datasets are chosen as the benchmark for our evaluation: 2D Hela dataset, PAP smear dataset, and Hep-2 cell image datasets. Since we mainly use CNNs as our feature descriptors, all images are taken in colors to preserve with as much information as possible.

3.1.1 2D Hela dataset

2D Hela (Boland & Murphy, 2001) is a dataset of fluorescence microscopy images of Hela cells that are stained with various organelle-specific fluorescent dyes. The dataset contains 10 classes, which are DNA (Nuclei), ER (Endoplasmic reticulum), Giantin (cis/medial Golgi), GPP130 (cis Golgi), Lamp2 (Lysosomes), Mitochondria, Nucleolin (Nucleoli), Actin, TfR (Endosomes), and Tubulin. The images have uniform size of 512×382 and typical images of each class are shown in Fig. 7.

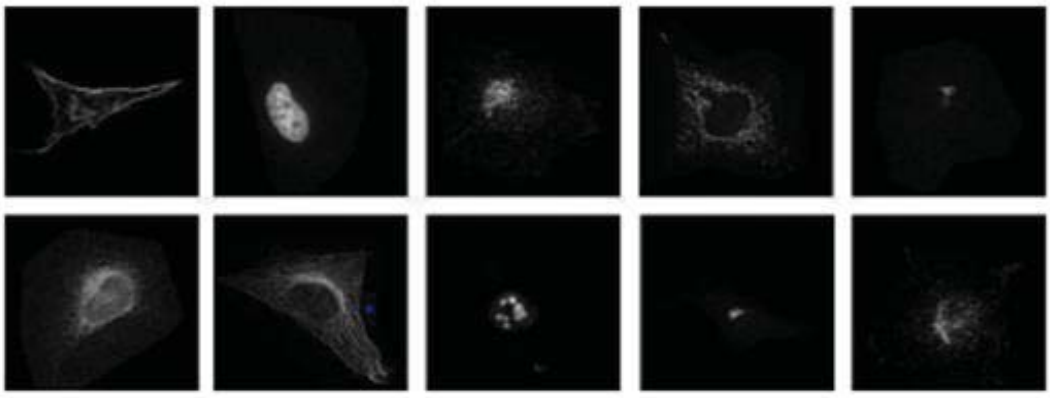


Figure 7. Typical images of 2D Hela dataset

3.1.2 PAP smear dataset

PAP smear dataset (Jantzen, et al., 2005) is a dataset which consists of 917 images unevenly distributed in seven different classes. It is published to the public to facilitate the detection of cervical cancer and serves as a benchmark for biomedical images classification. The image size ranges from 45×43 to 768×284 and representative samples are shown in Fig. 8.

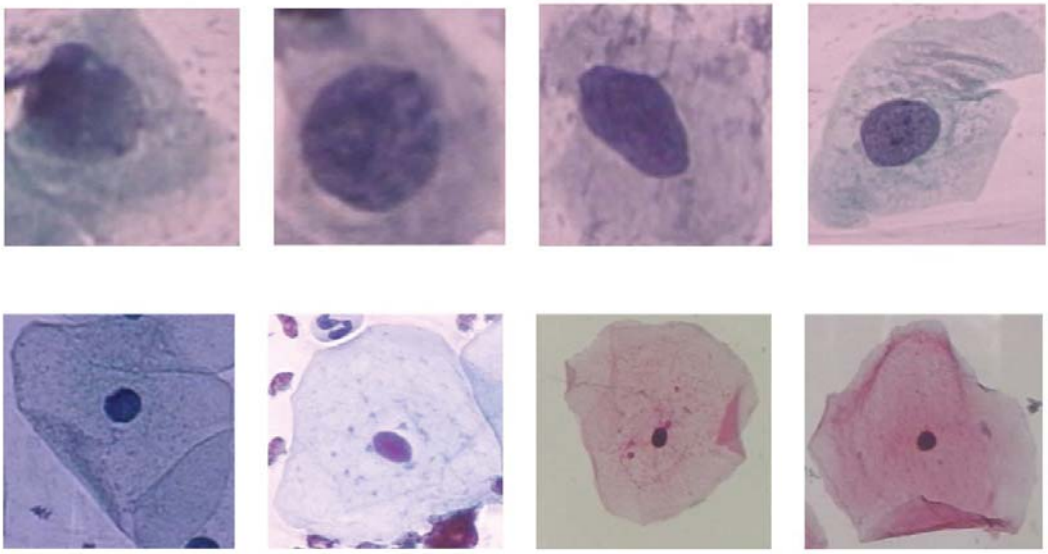


Figure 8. Typical images of PAP smear dataset in two categories: abnormal type (top row) and normal type (bottom row).

3.1.3 Hep-2 cell image dataset

The Human Epithelial type-2 cells or Hep-2 cells dataset (Foggia, et al., 2013) contains indirect immunofluorescence (IIF) images which can be applied to assist detection of autoimmune disease by searching for abnormal antibodies in the patient serum. The Hep-2 cell dataset is used by the contest of the International Conference on Pattern Recognition (ICPR) and we also use it as a benchmark to testify our ensembled model. The image size ranges from 33 × 38 to 396 × 295 and some typical images in the dataset are shown in Fig. 9.

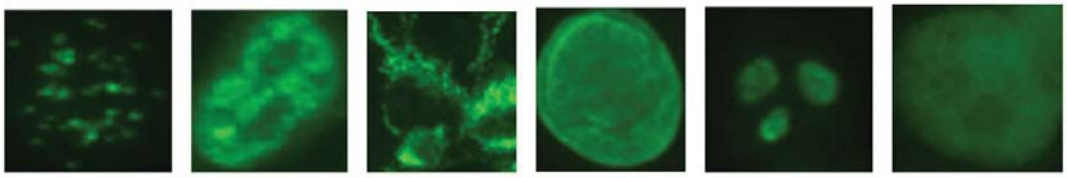


Figure 9. Typical images of Hep-2 cell image dataset

3.2 Experimental Setting

For each experiment, following (Lin et al., 2018), 30 independent trials are performed to obtain the average performance. The average accuracy and standard deviation for the testing set are then calculated accordingly. The accuracy a is calculated as follow:

$$a = \frac{m_0}{m} \times 100\%$$

in which m_0 is the number of correctly classified images and m is the total number of the testing images.

For the image pre-processing, since the raw images vary a lot in terms of size ranging from 33×38 to 512×382, we resize all images to 256×256 in order to fit them into the pre-trained model. Batch normalization is also performed based on the discussion in Section 2.3, while no data augmentation is carried out in the current setting except for the method presented in (Kumar, et al., 2017), to be discussed later. The fully connected layer of Inception-v3, ResNet152, Inception-ResNet-v2, and the concatenated feature map are 2048, 2048, 1536 and 5632, respectively. In each trial, we randomly choose 20% of the dataset as the testing set. For the remaining 80% of the dataset, we call it "training plus validation set" and randomly divide it to 4 equal folds to perform four-fold cross-validation to find the optimal hyperparameter using grid search. In each round of cross-validation, we choose a fold (which contains 20% of the original dataset) as the validation set and the other three folds (which contains 60% of the original dataset) as the training set. We then train the network on the training set, evaluate the performance on the validation set, and record the result. We repeatedly do that 4 times in order to iterate over the whole "training plus validation set". Finally, the average validation accuracy is taken as the baseline for choosing the best hyper-parameter setting. After determining the optimal set of the hyperparameters, the "actual" training is executed. The model is trained with the training and validation set using that hyperparameter setting and evaluated on the untouched test set.

Early stopping is used in the experiment to avoid overfitting. In the hyperparameter tuning step, validation-based early stopping is applied: the training will be stopped after 500 iterations if there is no improvement on validation accuracy. In the final training step, loss-based early stopping is applied: the training will be stopped if the training loss not decreasing after 500 iterations. The weights in the classification layers are randomly initialized by Gaussian distribution with zero mean and a standard deviation of 0.001. Exponential decay is used to calculate the adaptive learning rate and the formula for the learning rate is shown as follows:

$$\alpha = \alpha_0 \times k^{t/T}$$

where α_0: the initial learning rate;

k: the learning rate decay in hyperparameter setting;

T: the total number of steps which is set to a fixed number 1000;

t: the current time step when the learning rate is adaptively updated;

α: the updated learning rate at the current time step.

More details of hyper-parameter settings for each network are presented in Table 1.

Table 1. Hyper-parameter settings for each dataset and each network.

Dataset	*Architecture*	*Number of hidden layer neuron*	*Batch size*	*Base learning rate*	*Learning rate decay*	*Dropout keep probability*
PAP smear	Transfer learning with Inception-v3	50	50	0.075	0.5	0.8
	Transfer learning with ResNet152	100	30	0.075	0.33	0.7
	Transfer learning with Inception-ResNet-v2	50	30	0.03	0.33	0.8
	Feature concatenation of Inception-v3 & ResNet152	100	50	0.1	0.66	0.8
	Feature concatenation of Inception-v3 & ResNet152 & Inception-ResNet-v2	100	50	0.075	0.5	0.8
Hela	Transfer learning with Inception-v3	50	50	0.075	0.5	0.6
	Transfer learning with ResNet152	50	50	0.1	0.5	0.7
	Transfer learning with Inception-ResNet-v2	50	100	0.1	0.33	0.6
	Feature concatenation of Inception-v3 & ResNet152	50	30	0.1	0.33	0.8
	Feature concatenation of Inception-v3 & ResNet152 & Inception-ResNet-v2	50	30	0.075	0.5	0.7
Hep	Transfer learning with Inception-v3	150	30	0.1	0.33	0.7
	Transfer learning with ResNet152	150	30	0.05	0.66	0.7
	Transfer learning with Inception-ResNet-v2	50	50	0.03	0.33	0.8
	Feature concatenation of Inception-v3 & ResNet152	100	30	0.05	0.5	0.7
	Feature concatenation of Inception-v3 & ResNet152 & Inception-ResNet-v2	200	30	0.1	0.66	0.7

3.3 Experimental Results

Firstly, we compare the proposed feature concatenation method and the feature concatenation and ensemble method with each individual CNN models. Since the classification accuracy of Inception-ResNet-v2 is slightly lower than that of other two CNN models, we also experiment on removing it from the feature concatenation and ensemble method in order to examine its contribution. Table 2 shows the average classification accuracy and the standard deviation of the compared methods, from which we can obtain several observations: i) The feature concatenation and ensemble method outperforms each individual CNN and the feature concatenation method for all three datasets in terms of accuracy, which demonstrates the effectiveness of our proposed method. ii) Individual networks tend to produce a good performance for one dataset but might perform poorly for another dataset. For example, single Inception-v3 beats the other two for 2D Hela dataset and Hep2 dataset, while it is about 1% lower than ResNet152 in accuracy for PAP smear dataset. This indicates that CNNs with different depths and structures could extract distinctive features which may fit better to different datasets. iii) The comparisons of two feature concatenation and ensemble experiments show that after removing the Inception-ResNet-v2, the classification accuracy drops by about 1.0%, 0.6%, and 0.2% for the 2D-Hela, PAP smear, and Hep dataset, respectively. These observations indicate that Inception-ResNet-v2 still makes some contribution to the overall classification performance and it is helpful to incorporate it provided that enough training time and computational resources are available.

From the experimental results and the observations discussed above, we conclude that an ensemble of a diverse family of learning models provides a more reliable result for different datasets. By bypassing the fine-tuning and refining the network with hidden layers added to the most recently pre-trained models, the proposed feature concatenation and ensemble method achieves satisfactory computational efficiency without compromising the accuracy.

Secondly, we compare the proposed feature concatenation and ensemble method with two state-of-the-art methods: (i) the spatial adjacent histogram based on adapted local binary patterns (SAHLBP) (Liu, et al., 2016) and (ii) a reject option based cascade structure of a support vector machine (SVM) with subspace analysis (Lin et al., 2018). Based on the original LBP, SAHLBP proposes a spatial adjacent histogram strategy with an adaptive neighbourhood radius assigned to each pixel. Three coding schemes are discussed to encode the micro-structures and local patterns for biomedical images (Liu, et al., 2016). The reject option based method improves the performance by taking two complementary hand-craft features, scale-invariant feature transform (SIFT) and speed-up robust feature (SURF) and by applying reject option to ensure a relatively high confidence score for each classification result (Lin et al., 2018).

Table 2. Comparison of the classification accuracies among each individual neural network, the proposed feature concatenation method and feature concatenation ensemble method, with and without Inception-Resnet-v2, respectively.

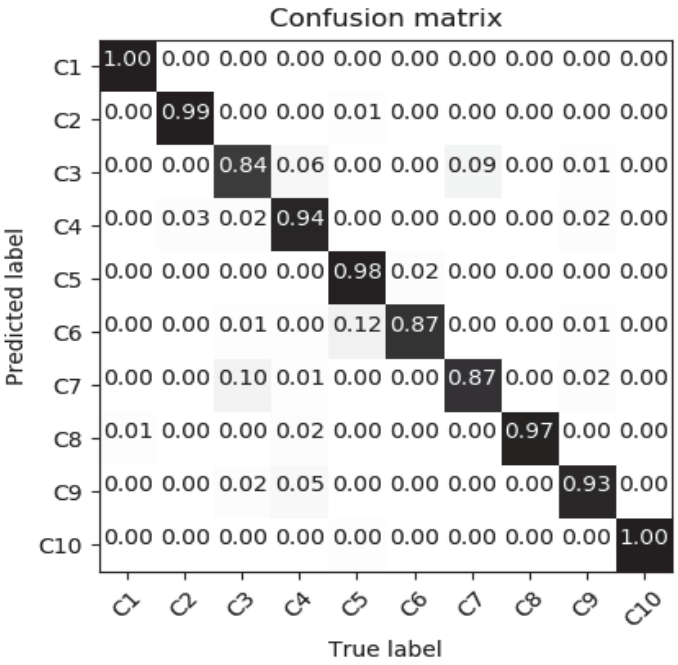
Methods	*2D-Hela*	*PAP smear*	*Hep*
Transfer learning with Inception-v3	91.42% ± 1.91%	89.66% ± 1.89%	92.95% ± 1.33%
Transfer learning with Resnet152	89.95% ± 2.30%	90.87% ± 1.48%	92.28% ± 1.59%
Transfer learning with Inception-Resnet-v2	91.79% ± 2.61%	89.25% ± 2.23%	89.45% ± 1.48%
Feature concatenation of Inception-v3 & Resnet152	92.06% ± 1.64%	92.01% ± 1.50%	94.10% ± 1.23%
Feature concatenation of Inception-v3 & Resnet152 & Inception-Resnet v2	92.57% ± 2.58%	92.63% ± 1.68%	94.86% ± 1.58%
Ensemble of Inception-v3 & Resnet152 & feature concatenation of the two	92.56% ± 1.92%	92.38% ± 1.28%	94.78% ± 1.05%
Ensemble of Inception-v3 & Resnet152 & Inception-Resnet-v2 & feature concatenation of the three	**93.51% ± 2.29%**	**93.04% ± 1.53%**	**94.98% ± 1.13%**

The average accuracy and standard deviation of the SAHLBP based method, reject option-based method, and the proposed feature concatenation and ensemble method are shown in Table 3. It is observed that the proposed method outperforms the reject option-based method for all three datasets by about 1.5% in accuracy. One possible reason is that unlike the reject option-based method that extracts hand-craft features like SIFT and SURF, the proposed feature concatenation and ensemble method takes advantages of multiple CNNs, which use the original three-channel color images without converting them into grayscale images. This exploits the discriminative information from three channels of color images for better classification. Another possible reason is that the state-of-the-art CNNs stand out with better generalization and robustness in performance compared with the traditional classification methods, e.g., SVM, when handling complex biomedical pattern recognition tasks. These findings indicate that ensemble of CNNs can produce better classification results for a specific domain without resorting to hand-craft features.

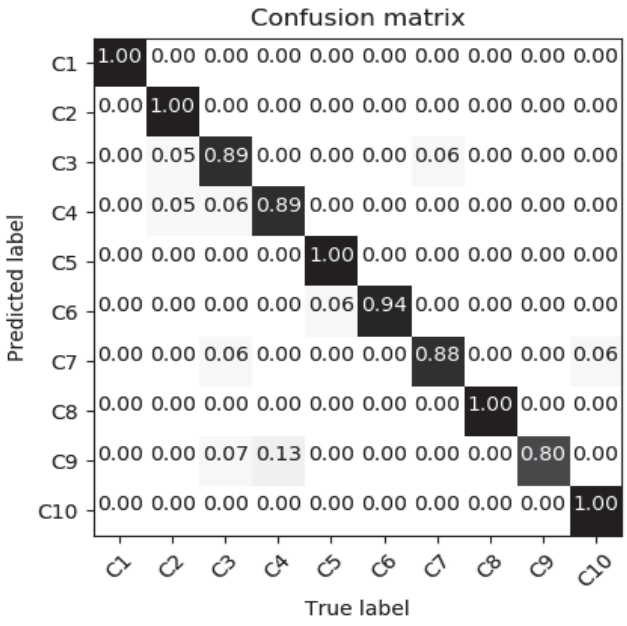
Table 3. Comparison of the classification accuracies among the SAHLBP based method, reject option-based method, and the proposed feature concatenation and ensemble method.

Methods	*2D Hela*	*PAP smear*	*Hep2*
SAHLBPT1 + SVM (Liu, et al., 2016)	89.68% ± 2.0%	86.69% ± 2.1%	91.89% ± 0.8%
SAHLBPT2 + SVM (Liu, et al., 2016)	87.29% ± 1.6%	84.03% ± 2.1%	88.59% ± 0.8%
SAHLBPT3 + SVM (Liu, et al., 2016)	90.06% ± 1.5%	88.03% ± 1.7%	91.86% ± 0.9%
Reject option based method (Lin et al., 2018)	92.96% ± 1.3%	90.96% ± 0.5%	92.97% ± 1.0%
Proposed feature concatenation and ensemble method	**93.51% ± 2.3%**	**93.04% ± 1.5%**	**94.98% ± 1.1%**

The confusion matrices for the proposed feature concatenation and ensemble method and the reject option based method (Lin et al., 2018) for the 2D Hela dataset and for the Hep-2 cell image dataset are shown in Figs. 10 and 11, respectively. Note that in general, the proposed feature concatenation and ensemble method performs better than the reject option based method, as can be observed from these two figures.

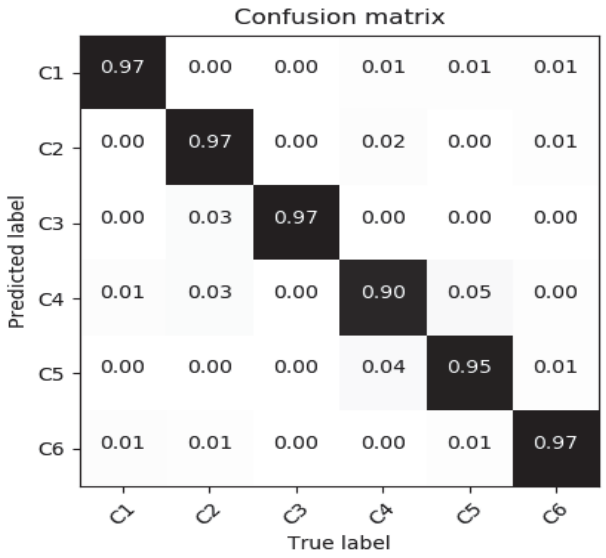


(a)

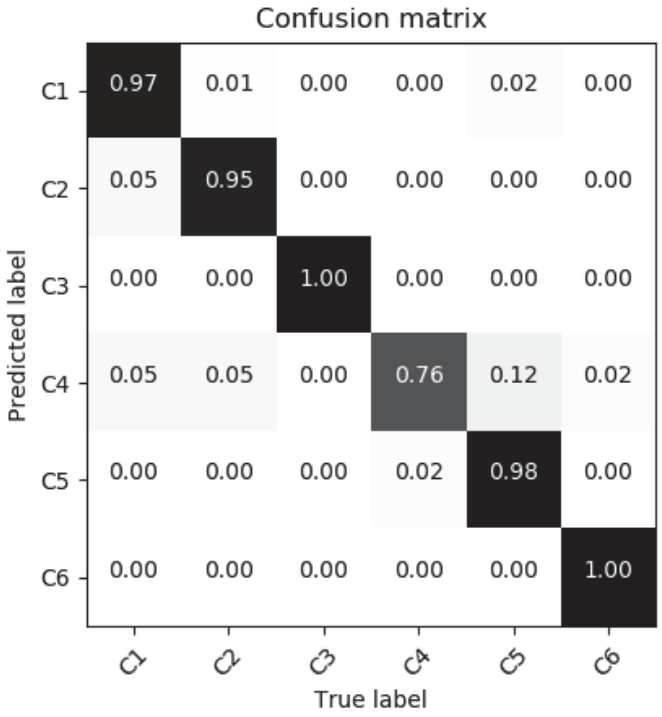


(b)

Figure 10. The confusion matrices for the 2D Hela dataset generated by (a) the proposed feature concatenation and ensemble method; (b) the rejection option based method (Lin et al., 2018). C1: Actin; C2: DNA; C3: Endosome; C4: Er; C5: Golgia; C6: Golgpp; C7: Lysosome; C8: Microtubules; C9: Mitochondria; C10: Nucleolus.



(a)



(b)

Figure 11. The confusion matrices for the Hep-2 cell image dataset generated by (a) the proposed feature concatenation and ensemble method; (b) the rejection option based method (Lin et al., 2018). C1: Centromere; C2: Coarse speckled; C3: Cytoplasmic; C4: Fine speckled; C5: Homogeneous; C6: Nucleolar.

Finally, we compare our proposed feature concatenation and ensemble method with the method presented in (Kumar, et al., 2017) which exploits multiple fine-tuned CNNs as feature extractors for the classification of medical images. We follow the ensemble design and re-do the fine-tuning part for both AlexNet and GoogleNet using our 2D-Hela dataset. Note that the images in our other two datasets, namely PAP smear and Hep, are too small to implement data augmention by cropping and flipping the center and four corners of each image, which is part of the method proposed in (Kumar, et al., 2017). We use the Stochastic Gradient Descent (SGD) as the optimizer for both GoogleNet and Alexnet. Regarding the parameter setting, we experiment on momentum = 0.9 as mentioned in (Kumar, et al., 2017) and momentum = 0, respectively. For the learning rate, we start from a large range of 1e-6 to 0.1 and find out that the range of [0.001 - 0.05] is suitable for the 2D-Hela dataset. Then we perform hyper parameter search for learning rate in [0.001-0.05] with step of 0.025 and introduce learning rate decay chosen in [0, 1e-6, 1e-3]. We split our dataset following our other experiments as stated in Section 3.2. Early stopping based on validation accuracy is introduced, i.e., the training will stop if the validation accuracy is not improved by more than 0.1% after 5 consecutive steps. The maximum number of epochs is set to 50, following (Kumar, et al., 2017). The final parameters chosen for AlexNet and GoogleNet are stated in Table 4.

Table 4. Parameters chosen for AlexNet and GoogleNet.

Network	Momentum	Learning Rate	Learning Rate Decay
AlexNet	0.9	0.0025	1e-06
GoogleNet.	0.9	0.001	1e-06

After fine-tuning, we also apply Principle Component Analysis (PCA) to reduce the dimensionality for efficient SVM classifier training. Instead of directly choosing principle components that keep 90% of the data variance, we experiment on the PCA dimensions which occupy 90% and 95% of the variance, respectively. Finally, the one-vs-one multi-class linear SVM classifiers are trained with the parameter C selected from the range [2×10^{-15}, 2×10^{-14}, ... , 2×10^{14}, 2×10^{15}] (Kumar, et al., 2017).

During the implementation, we notice that the classification accuracy of the method of (Kumar, et al., 2017) is much lower than that of our proposed method. We investigate the implementation of their method and suspect that it is the image pre-processing step that might impair the result. As shown on the left side of Fig. 12, if we strictly follow (Kumar, et al., 2017) and perform data augmention by cropping and flipping the center and four corners of each image, some important information might be lost. This is due to the fact that useful information may not be contained in the whole image and hence some of the sub-images contain little or no information about the targets in the images of our dataset. With this observation, we re-implement the method of (Kumar, et al., 2017) again, but without cropping and flipping the center and four corners of each image this time.

The results for both experiments, with and without image cropping and flipping the center, are shown in Table 5. It is noted that although the method in (Kumar, et al., 2017) produces better classification accuracy for the 2D-Hela dataset without cropping and flipping, our proposed method still achieves an accuracy gain by about 1.5% compared to the method in (Kumar, et al., 2017). The main reason for such performance gain comes from our exploitation of more advanced CNN architectures like ResNet and by adding one hidden layer before the last soft-max layer in our proposed method.

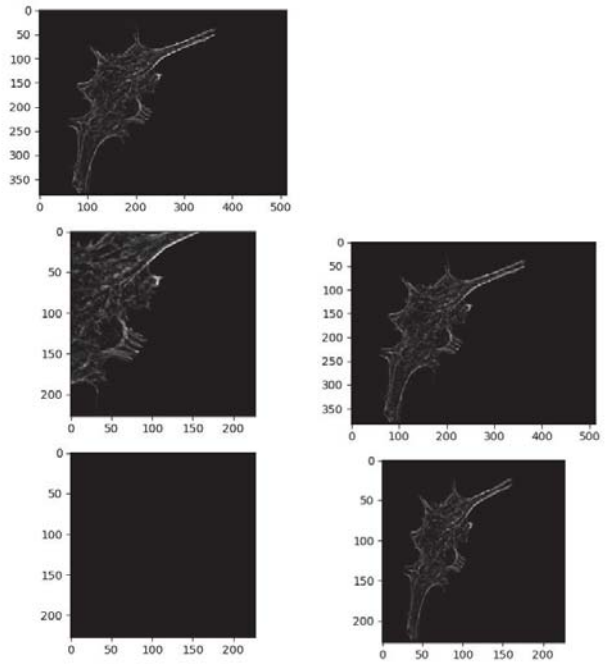


Figure 12. Demonstration of information loss resulting from a) cropping method compared to b) normal resizing method. For a): original image (top), cropped center image (middle), cropped bottom image (bottom); for b) original image (top), resized image (bottom).

Table 5. Comparison of the classification accuracies for Hela dataset among different networks used in the CNN-related method (Kumar, et al., 2017), with and without image cropping and flipping respectively, and the proposed feature concatenation and ensemble method.

Methods	*With cropping and flipping?*	PCA 95%	PCA 90%
Finetuned AlexNet + softmax	Yes	70.15 ± 4.43	70.15 ± 4.43
Finetuned AlexNet + SVM	Yes	90.23 ± 2.26	89.33 ± 2.29
Finetuned GoogleNet + softmax	Yes	76.20 ± 4.57	76.20 ± 4.57
Finetuned GoogleNet + SVM	Yes	91.29 ± 1.80	90.67 ± 2.46
Concatenation of Finetuned AlexNet & Finetuned GoogleNet + SVM	Yes	91.98 ± 1.15	91.44 ± 1.36
Ensemble of five (Kumar, et al., 2017)	Yes	91.54 ± 2.05	90.92 ± 1.57
Finetuned AlexNet + softmax	No	81.71 ± 5.01	81.71 ± 5.01
Finetuned AlexNet + SVM	No	84.87 ± 4.11	81.23 ± 4.66
Finetuned GoogleNet + softmax	No	92.02 ± 1.54	92.02 ± 1.54
Finetuned GoogleNet + SVM	No	91.75 ± 2.08	91.10 ± 2.52
Concatenation of Finetuned AlexNet & Finetuned GoogleNet + SVM	No	91.64 ± 1.92	91.62 ± 1.85
Ensemble of five (Kumar, et al., 2017)	No	92.02 ± 1.87	91.68 ± 1.75
Proposed feature concatenation and ensemble method		**93.51 ± 2.29**	

4. Conclusion

In this paper, we have proposed a feature concatenation method for biomedical image classification based on three individual models, Inception-v3, ResNet152, Inception-ResNet-v2. To further improve the classification performance, we also proposed the feature concatenation and ensemble method consisting of four individual models, Inception-v3, ResNet152, Inception-ResNet-v2. and the feature concatenated CNN model. In the experiments, three benchmark datasets are used for testing, each consisting of different patterns and features. The results show that the feature concatenation and ensemble method generally outperforms each of the individual network, including the feature concatenation method, as well as several competing methods, both with or without CNNs, in terms of classification accuracy. The main contribution of this work is that transfer learning, feature concatenation and ensemble learning are integrated so that satisfactory results can be obtained for biomedical image classification even though the dataset is relatively small for the specific task. In particular, by adding one hidden layer before the last soft-max layer, we avoid the fine-tuning procedure.

As a result, the proposed feature concatenation and ensemble method manages to make a balance between accurate classification and computational efficiency, which is essential for automated biomedical image classification and hence in the long run contributes to the establishment of a smart healthcare system in real life.

Acknowledgements

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on Campus (URECA) program.

References:

Ambikapathy, B., & Krishnamurthy, K. (2018). Analysis of electromyograms recorded using invasive and noninvasive electrodes: a study based on entropy and Lyapunov exponents estimated using artificial neural networks. *Journal of Ambient Intelligence and Humanized Computing*, 1-9.

Ashtarian, H., Mirzabeigi, E., Mahmoodi, E., & Khezeli, M. (2017). Knowledge about Cervical Cancer and Pap Smear and the Factors Influencing the Pap test Screening among Women. *International Journal of Community Based Nursing and Midwifery, 5*(2), 188-195.

Boland, M. V., & Murphy, R. F. (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics, 17*(12), 1213-1223.

Ciresan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks . *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 411-418). Nagoya, Japan: Springer, Berlin, Heidelberg.

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. *Neural Information Processing Systems 2012*, *2*, pp. 2843-2851. Lake Tahoe, Nevada, USA.

Dietterich, T. G. (1997). Machine Learning Research: Four Current Directions. *AI Magazine, 18*(4), 7-136.

Duneja, A., Puyalnithi, T., Vankadara, M. V., & Chilamkurti, N. (2018). Analysis of inter-concept dependencies in disease diagnostic cognitive maps using recurrent neural network and genetic algorithms in time series clinical data for targeted treatment. *Journal of Ambient Intelligence and Humanized Computing*, 1-9.

Esteva, A., Kuprel, b., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

Foggia, P., Percannella, G., Soda, P., & Vento, M. (2013). Benchmarking Hep-2 cells classification methods. *IEEE Transactions on Medical Imaging, 32*(10), 1878-1889.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). Las Vegas, Nevada: IEEE.

Jantzen, J., Norup, J., Dounias, G., & Bjerregaard, B. (2005). PAP-smear benchmark data for pattern classification. *Nature inspired Smart Information Systems* (pp. 1-9). Albufeira, Portugal: NiSIS.

Jeon, G. (2017). Computational intelligence approach for medical images by suppressing noise. *Journal of Ambient Intelligence and Humanized Computing*, 1-11.

Ju, C., Bibaut, A., & van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics, 45*(15), 2800-2818.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., . . . Glocker, B. (2017). Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis, 36*, 61-78.

Kawahara, J., & Hamrneh, G. (2016). Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers. *Machine Learning in Medical Imaging MLMI 2016.* (pp. 164-171). Athens, Greece: Springer, Cham.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. *Proceedings of the 7th International Conference on Neural Information Processing Systems* (pp. 231-238). MIT press.

Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2017). An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. *IEEE Journal of Biomedical and Health Informatics, 21*(1), 31-40.

Lin, D., Sun, L., Toh, K.-A., Zhang, J., & Lin, Z. (2018). Biomedical image classification based on a cascade of an SVM with a reject option and subspace analysis. *Computers in Biology and Medicine, 96*, 128-140.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., . . . Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88.

Liu, D., Wang, S., Huang, D., Deng, G., Zeng, F., & Chen, H. (2016). Medical image classification using spatial adjacent histogram based on adaptive local binary patterns. *Computers in Biology and Medicine, 72*, 185-200.

Nguyen, L. D., Lin, D., Lin, Z., & Cao, J. (2018). Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. *2018 IEEE International Symposium on Circuits and Systems (ISCAS).* Florence, Italy, May, 2018.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Advancement of Artificial Inteligence.* San Francisco, California, USA.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Computer Vision and Pattern Recognition 2016*, (pp. 2818-2826). Las Vegas, Nevada.

Zheng, L., Zhao, Y., Wang, S., Wang, J., & Tian, Q. (2016). *Good Practice in CNN Feature Transfer.* ArXiv.