Comparison of Deep Learning-Based Recognition Techniques for Medical and Biomedical Images

Majtner, Tomas; S. Nadimi, Esmaeil

Published in:
International Conference on Computer Analysis of Images and Patterns

DOI:
10.1007/978-3-030-29888-3_40

Publication date:
2019

Document version:
Accepted manuscript

Citation for pulished version (APA):
Majtner, T., & S. Nadimi, E. (2019). Comparison of Deep Learning-Based Recognition Techniques for Medical and Biomedical Images. In M. Vento, & G. Percannella (Eds.), *International Conference on Computer Analysis of Images and Patterns: CAIP 2019: Computer Analysis of Images and Patterns* (pp. 492-504). Springer. https://doi.org/10.1007/978-3-030-29888-3_40

Comparison of Deep Learning-Based Recognition Techniques for Medical and Biomedical Images

Tomáš Majtner and Esmaeil S. Nadimi

Group of Machine Learning and AI, The Maersk Mc-Kinney Moller Institute,
University of Southern Denmark, Odense, Denmark
Contact email: tomaj@mmmi.sdu.dk

Abstract. The recognition and classification of medical and biomedical images typically suffer from the problem of a low number of annotated samples. This comes along with the problem of efficient training of the current deep learning frameworks. Therefore, many researchers opt for various techniques which could substitute the traditional training of convolutional neural networks (CNN) from scratch. In this article, we are comparing multiple of these methods, including transfer learning and using the CNNs as feature extractors. The paper contains results on two datasets with different modalities and three different CNN architectures. We demonstrate the high effectiveness of transfer learning and suggest that, in some cases, it is worth retraining more layers at the end of the network for achieving higher accuracy.

Keywords: Image recognition · GoogLeNet · VGG-16 · ResNet-50 · Transfer learning · Polyp detection · HEp-2 image classification.

1 Introduction

The recent progress in machine learning algorithms has a significant impact on all domains where image analysis is present. New recognition methods, typically based on convolutional neural networks (CNN) and so-called *deep learning* approach, are presented in the literature with high frequency. The domains of medical and biomedical image analysis are typical examples. Only since last year, we can find a guide to deep learning in healthcare [9], applications in biomedicine [4, 33], medical image analysis [18], biomedical data science [2], and many others.

Image recognition methods and applications based on deep learning in these domains are easily accessible, and they are still attracting a lot of attention. However, many questions regarding their optimal usage remain open. One of them, which is of our interest in this paper, relates to the training process. To design and train any accurate deep learning recognition framework, a large amount of training samples is required. Depending on the network architecture, we need a certain minimal number of input images to set all the weights in network layers for precise categorisation of test images.

In a real-world image recognition scenario, networks are typically trained on ImageNet [7], where millions of images are hand-annotated. However, in medical

and biomedical domains, we must deal with a lack of annotated images. There are some specific methods based on the reduction of annotation effort by making judicious suggestions on the most effective annotation areas [34], or based on image augmentation [28], or generative adversarial networks (GANs) [26] that could be used to increase the size of our datasets. However, even these methods will not properly simulate the quality and variability of large real-world datasets.

Therefore, many researchers in medical and biomedical domains are not training networks from scratch, but they rather use a transfer learning approach [35]. In this scenario, networks pretrained on real-world images are used, with only the last layers of the network modified to classify new images to desired categories. All the weights from the original network are simply copied. However, one can also opt for retraining more layers at the end of the network and transfer/copy only those weights that correspond to the first layers.

Another popular option is to use the CNN as a feature extractor [19], where values from the last layers, typically fully-connected ones, are used together with an external classifier like the discriminant analysis (LDA) or the k-nearest neighbour classifier (k-NN). This approach often leads to overlooking the curse of dimensionality [17, 22].

Because of the variability of above-mentioned methods, we decided to compare all of these options for two particular datasets. The first one is medical and consists of images from colon capsule endoscopy, where we aim to detect polyps. The second one is biomedical and contains images of human epithelial (HEp-2) cells. These images are used in indirect immunofluorescence tests to detect autoimmune diseases. Even though both datasets are visually very different (compare Fig. 1 vs. Fig. 2) and come from different domains, in both cases the sample categorisation is currently mostly done by humans. This approach is often subjective, too dependent on the experience of the expert, and with the increasing number of new samples also expensive. Thus, computer-aided systems in both cases aim to assist doctors with correct and fast diagnosis.

Our motivation is to provide a fair comparison for both datasets to see if we can find any patterns which could be later generalised also for other medical and biomedical datasets. For this purpose, we are using three different network architectures, namely GoogLeNet [31], VGG-16 [30], and ResNet-50 [12].

In the next section, we will shortly describe the recent related work in the field and recent progress in polyp detection and HEp-2 image recognition. Subsequently, we will introduce our datasets, pre-processing methods, and an overview of all compared configurations. At the end of the article, we present the results together with a detailed discussion, conclusions, and suggested future work.

2 Related Work

The comparison between transfer learning and training the network from scratch for medical image analysis was covered by Tajbakhsh et al. [32]. Based on their experiments, transfer learning is more robust. A similar comparison was made by Shin et al. [29] who focused on thoraco-abdominal lymph node detection and

interstitial lung disease, and they also came to a similar conclusion. The study on the dependence of the detection performance on the extent of transfer for kidney images was done by Ravishankar et al. [27].

In the domain of polyp detection, Mamonov et al. [23] presented an automated method with 81 % sensitivity per polyp at a specificity level of 90 %. A review of polyp detection and segmentation from video capsule endoscopy was published in 2017 [24]. Many successful methods can be found that are still relying on the traditional recognition approach that is based, for example, on the Gabor texture features and the K-means clustering [14] or the scale-invariant feature transform and the complete local binary pattern [37]. However, the translation of information technology research to clinical practice is still limited [15].

The recent progress in the HEp-2 image analysis has been covered by a special issue of Pattern Recognition Letters [11]. Novel techniques including those examining the role of Gaussian Scale Space theory as a pre-processing approach [25], a superpixel-based classification method calculating the sparse codes of image patches [8], a multi-process system based on an ensemble of 15 support vector machines [5], and many others were introduced. More recently, Gao et al. [10] analysed the impact of hyper-parameter settings of proposed fully-connected CNN on the classification accuracy. The influence of several pre-processing techniques on HEp-2 image classification was studied by Bayramoglu et al. [3].

Our contribution differs from the previous similar comparative studies by direct evaluation of multiple transfer learning options together with the training from scratch strategy and with the usage of CNNs as feature extractors. We are also offering a comparison between two different image modalities trained on the same conditions using three different network configurations. The previous studies usually focused only on a specific domain and/or particular network configuration and/or comparison of the particular two training strategies.

3 Datasets

In this paper, we are using two different datasets. The POLYP DATASET consists of 854 individual frames that were acquired using PillCam^{TM} COLON 2 System, extracted from the videos, and annotated at Odense University Hospital. This dataset consists of 613 polyp images and 241 non-polyps. Because the original images contain text information near the corners, we used a Chan-Vese segmentation via graph cuts [6] to extract the binary mask and segment the relevant information from the image (see Fig.1 for illustration).

We randomly split the original dataset into 70 % for training, 10 % for validation, and 20 % for testing. Some researchers rather opt for N-fold cross-validation over all available images, especially when the dataset is small. However, this approach leads to biased results [1], where the performance tends to drop significantly when the algorithm is applied to new, previously unseen data. Therefore, we use a separate validation part to evaluate the performance during the training of deep learning and independent testing part to report the final performance.

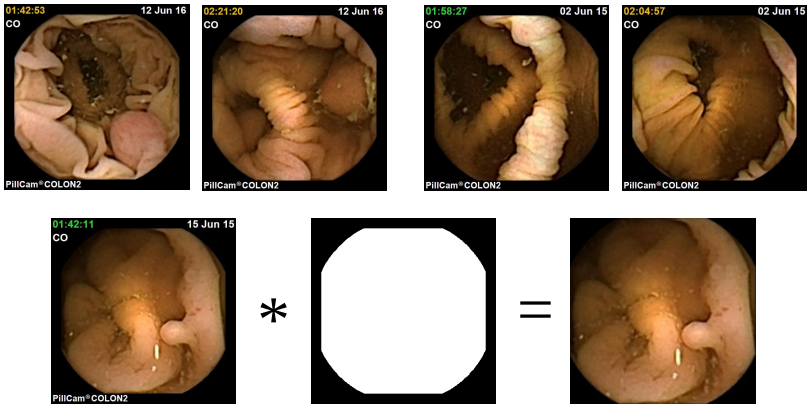

Fig. 1: **The top row**: examples of the images from the POLYP DATASET. First two from the left contain polyps, while the remaining two do not contain polyps. **The bottom row**: illustration of the pre-processing for one particular sample.

Because the number of samples in our dataset is small, we decided to augment the training and validation part of the dataset. The most natural technique to augment these images is to use image rotation around the image centre and the flipping or reversing operation, where the image is generated as mirror-reversal of an original one across a horizontal axis. First, we rotated each training and validation image by 90°, 180°, and 270°. Since the dataset is unbalanced, we decided to further augment each image from the non-polyp class of training dataset by 30° and 60°. Together with the flipping operation, this augmentation leads to the significant increase of dataset images. The total number of samples in the POLYP DATASET after above specified augmentation of training and validation dataset is summarized in Table 1.

Table 1: The total number of images in the POLYP DATASET after augmentation.

	Polyp	Non-polyp	Total
Training after augm.	3,424	4,032	7,456
Validation after augm.	488	192	680
Testing	124	49	173
Total	4,036	4,273	8,309

The HEP-2 DATASET is a publicly available dataset that was used for benchmarking [13]. The original dataset contains 13,596 pre-segmented and annotated cell images with their ground truth classes. The specimens, one for each patient serum, were automatically photographed using a monochrome high dynamic range cooled microscopy camera. Since the brightness and contrast of the images vary a lot, we performed intensity adjustment by linear stretching, where 1 % of the pixels are saturated at the low and the high end of the intensity range in order to maximize the contrast. This dataset is divided into six categories:

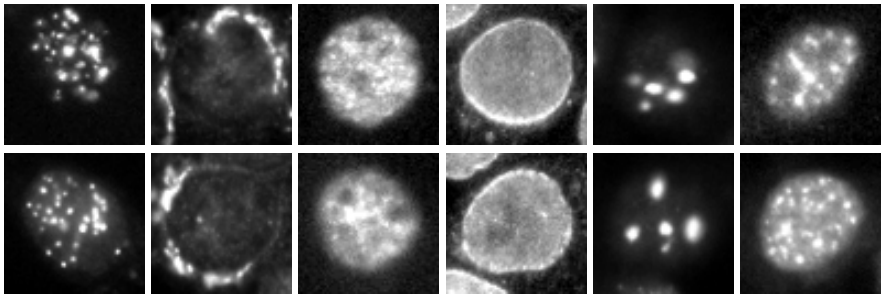

Fig. 2: Examples of images from the HEp-2 DATASET. Each column represents a different image class in order: Ce, Go, Ho, Nm, Nu, Sp.

Centromere (Ce), Golgi (Go), Homogeneous (Ho), Nuclear Membrane (Nm), Nucleolar (Nu), and Speckled (Sp). See Fig.2 for illustration.

Also here we used the same random split into 70 % for training, 10 % for validation, and 20 % for testing. Since we have more samples here, we decided to further augment only training part of this dataset. We used the rotation around the image centre by 90°, 180°, and 270° plus the flipping operation, as for the previous dataset. The HEp-2 DATASET is also unbalanced with one class (Golgi) having $3-4\times$ lower number of images than the remaining five classes. Therefore, we additionally rotated each Golgi image by angles of size $23^\circ \times i$, where $i \in \{1, 2, 3\}$, where these rotated images were cropped to the size of the largest rectangle within the input image. After this step, Golgi class has similar size than the remaining classes. The total number of samples in each class after the augmentation is summarised in Table 2.

Table 2: The total number of images in the HEp-2 DATASET after augmentation.

	Ce	Go	Ho	Nm	Nu	Sp	Total
Training after augm.	15,344	16,192	13,960	12,368	14,552	15,848	88,264
Validation	274	72	249	220	259	283	1,357
Testing	549	146	500	442	520	567	2,724
Total	16,167	16,410	14,709	13,030	15,331	16,698	92,345

4 Overview of Compared Configurations

As it was mentioned before, we are considering three different approaches for image recognition that are connected to deep leaning. The first one is based on extracting features and using external classifier. This methodology is common for medical image analysis [16, 19, 21], and its main advantage is that it does not require high computational resources. The idea is that we only download the pretrained model of the network, typically fine-tuned for the real-world images, and use it as a feature extractor for each image from our dataset. These features are subsequently used as an input to a classifier, as it is depicted in Fig. 3.

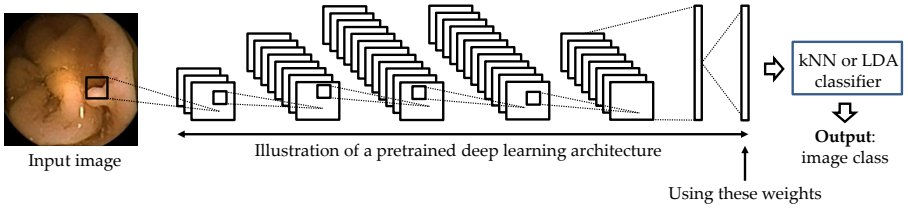

Fig. 3: Illustration of the principle, where a deep learning architecture is used as a feature extractor.

We employed two different classifiers, namely the k-nearest neighbours classifier (k-NN) and the discriminant analysis (LDA). Features extracted from the training dataset were used to train the classifiers and features from validation dataset to set the hyper-parameters. For better comparison, we extracted features from two different layers from each of the pretrained networks. Table 3 offers the summary of used layers for each network architecture together with the size of the feature vector length.

Table 3: Names and sizes of layers that are used to extract features for each architecture.

GoogLeNet	pool5-7x7_s1 (1024)	VGG-16	fc7 (4096)	ResNet-50	avg_pool (2048)
	loss3-classifier (1000)		fc8 (1000)		fc1000 (1000)

Because the dimensionality of extracted features, specified in brackets in Table 3, is very high in comparison with our test dataset, we employed principal component analysis (PCA) to reduce it. Using this approach, we were able to decrease the number of used features down to 20 for k-NN classifier with achieving the same classification accuracy. For LDA classifier, we reduced the dimensionality to approximately 200, while keeping the same accuracy. The exact numbers will be presented in the next section.

The second approach for image recognition used in this study is based on training the deep learning architecture from scratch. In this scenario, we consider only the architecture, but we ignore the pretrained weights. As it was pointed out in previous studies [36], this approach is common for computer vision, where large datasets are available. For the medical and biomedical image processing, its usability and effectiveness are, however, limited [32]. Often the dataset is not large enough to train these networks and researchers tend to design smaller architecture. In this study, we are comparing only three predefined architectures, we are not attempting to design new ones. Therefore, we will only report the results on those configurations, where the training was possible due to the dataset size restrictions.

The last approach is based on transfer learning [35], which is generally very popular for medical and biomedical image recognition [20, 29, 32]. Here, all three network architectures were pretrained on ImageNet [7], and we re-

placed/retrained their last layers to classify images directly to our two categories for the POLYP DATASET and six categories for the HEP-2 DATASET.

We also decided to test how far we can go with retraining the pretrained layers. In traditional transfer learning, we typically only replace layers responsible for classification to specified categories. However, researchers could retrain also some convolutional layers (CONV) to better capture the specific fine characteristics of their dataset. In theory, this should be important especially when we are transferring the knowledge between image modalities. Specifically in our case, in GoogLeNet we replaced last 8 layers (incl. 1 CONV), last 19 layers (incl. 6 CONV), and last 33 layers (incl. 12 CONV); in VGG-16 we replaced last 16 layers (incl. 3 CONV) and last 23 layers (incl. 6 CONV); in ResNet-50 we replaced last 8 layers (incl. 1 CONV), last 14 layers (incl. 3 CONV), and last 30 layers (incl. 8 CONV).

All images were resized to 224×224 to fit the input size of networks. The experiments were performed using MATLAB R2018b, and hyper-parameters were fine-tuned using validation dataset. For the training of each CNN, either from scratch or via transfer learning, we utilised stochastic gradient descent with momentum optimiser. The HEP-2 DATASET was trained for 30 epochs with a learning rate of 10^{-3} and a mini-batch size of 32 images. Since the POLYP DATASET is much smaller, we used a mini-batch size of 4 images, and after we trained them for 30 epochs with a learning rate of 10^{-3}, we added 30 more epochs with a learning rate of 10^{-4}. This decrease in learning rate had a positive effect on convergence towards minimum validation loss. We did not observe this effect for the HEP-2 DATASET. We cannot provide here all the graphs from the training process, but the illustration of the loss development for transfer learning using GoogLeNet is depicted in Fig. 4.

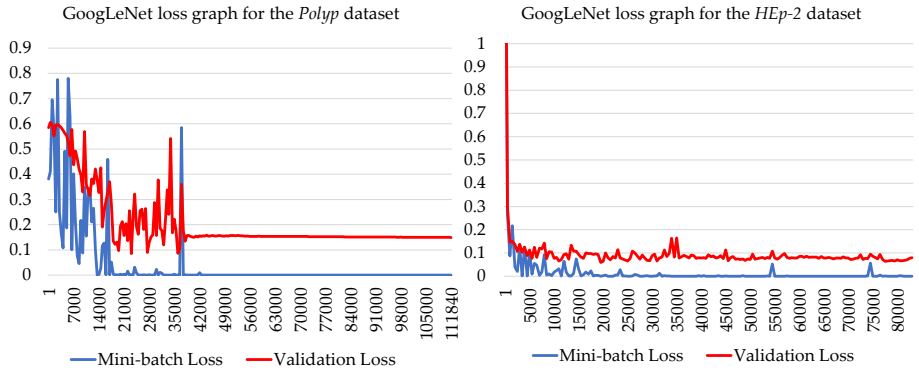

Fig. 4: Graphs depicting the loss for transfer learning using GoogLeNet on both datasets. The x-axis shows the number of iterations.

5 Results and Discussion

The evaluation of classification performance is done using various metrics. In the Polyp dataset, we have a binary classification problem, where the most commonly used metrics include classification accuracy (ACC), sensitivity (TPR), and specificity (SPC). They are defined as follows:

$$\text{ACC} = \frac{\text{true positives} + \text{true negatives}}{\text{number of samples}}, \tag{1}$$

$$\text{TPR} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \tag{2}$$

$$\text{SPC} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}. \tag{3}$$

For the evaluation of the HEp-2 dataset, we use the classification accuracy (ACC), which is defined here as the overall correct classification rate of all images. The second evaluation metric is the mean class accuracy (MCA), which is defined as

$$MCA = \frac{1}{K}\sum_{k=1}^{K} CCR_k \tag{4}$$

where CCR_k is the classification accuracy of a particular cell class k and K is the number of cell classes.

The comparison of all tested variants for the Polyp dataset is summarized in Table 4 and for the HEp-2 dataset in Table 5. In all cases, extracting the features from deep learning and using them with external classifier performed worse when compared to methods based on transfer learning. As it was expected, even the training from scratch was not performing well. In many cases, it was not possible to train the network at all, due to the limited number of samples but even when the training was successful, the performance was low.

The most interesting result, however, is the comparison between transfer learning strategies. From the results, we can see that retraining last layers can have a positive effect on the final performance of the recognition framework. This was observed especially for GoogLeNet and ResNet-50 architectures, and we see this effect on both datasets, which means that it is not specific for one domain or one type of images. We can also observe the trend that with retraining more layers, the performance typically starts to decline, which could be ascribed to low number of samples.

The results on the Polyp dataset are not comparable with the literature since the dataset is private, but the comparison of our top performing results for the HEp-2 dataset with the methods from the literature is presented in Table 6. We included only those results that are using a similar split technique for training and testing dataset. The method presented by Shen et al. [28] proposed a deep

Table 4: The summary of results for the POLYP DATASET. The number of nearest neighbours in k-NN classifier is specified in brackets. In transfer learning, *end* represents the retraining of layers responsible for classifying, while the *lastX* variants represent the retraining of last X layers. Presented values are in %.

GoogLeNet									
	Features				Scratch	Transfer learning			
	pool5-7x7_s1		loss3-classifier						
	k-NN(10)	LDA	k-NN(11)	LDA		end	last 8	last 19	last 33
ACC	77.46	75.72	75.14	76.88	–	91.33	**93.06**	89.02	90.75
TPR	91.94	87.10	93.55	87.90	–	96.77	**98.39**	96.77	95.16
SPC	40.82	46.94	28.57	48.98	–	77.55	**79.59**	69.39	79.59

VGG-16								
	Features				Scratch	Transfer learning		
	fc7		fc8					
	k-NN(10)	LDA	k-NN(10)	LDA		end	last 16	last 23
ACC	76.30	79.19	76.88	78.03	–	**94.22**	90.17	89.59
TPR	91.94	87.10	92.74	90.32	–	**96.77**	95.97	95.97
SPC	36.73	59.18	36.73	46.94	–	**87.76**	75.51	73.47

ResNet-50									
	Features				Scratch	Transfer learning			
	avg_pool		fc1000						
	k-NN(11)	LDA	k-NN(10)	LDA		end	last 8	last 14	last 30
ACC	78.61	84.39	79.19	84.97	74.57	**94.22**	90.75	90.75	80.34
TPR	90.32	87.10	91.94	88.71	83.87	**97.58**	91.93	94.35	87.09
SPC	48.98	77.55	46.94	75.51	51.02	85.71	**87.76**	81.63	63.26

cross residual network (DCRNet) for HEp-2 cell classification, and it was the winner of the most recent HEp-2 image recognition contest.

6 Conclusion

In this article, we presented a comparison of three different techniques for image recognition, where all of them are based on deep learning. We used three different network configurations to capture the effect of architecture choice on final results, and we tested two different image modalities, which currently share the same problem with manual annotation. Our results demonstrate the high performance of our solutions, which are demonstrated especially on the HEP-2 DATASET. We also observed that the retraining of the last layers can be very efficient in terms of the final performance. Moreover, this conclusion was observed on both datasets, which means that it is not domain dependent.

The future work should be concentrated on the improvement of performance on the POLYP DATASET. It is always challenging to build a framework for a domain with very few samples, and therefore we would like to explore more the possibilities of increasing the number of samples by using generative adversarial networks.

Table 5: The summary of results for the HEp-2 dataset. The number of nearest neighbours in k-NN classifier is specified in brackets. In transfer learning, *end* represents the retraining of layers responsible for classifying, while the *lastX* variants represent the retraining of last X layers. Presented values are in %.

GoogLeNet									
	Features				Scratch	Transfer learning			
	pool5-7x7_s1		loss3-classifier						
	k-NN(10)	LDA	k-NN(12)	LDA		end	last 8	last 19	last 33
ACC	88.77	90.05	87.26	86.78	–	**98.53**	98.49	98.38	98.05
MCA	87.67	89.34	85.74	85.18	–	**98.64**	98.62	98.44	98.21

VGG-16								
	Features				Scratch	Transfer learning		
	fc7		fc8					
	k-NN(8)	LDA	k-NN(15)	LDA		end	last 16	last 23
ACC	88.91	91.37	87.11	86.23	94.38	**98.16**	97.61	97.54
MCA	87.06	90.53	84.84	84.24	94.89	**98.22**	97.53	97.62

ResNet-50									
	Features				Scratch	Transfer learning			
	avg_pool		fc1000						
	k-NN(10)	LDA	k-NN(9)	LDA		end	last 8	last 14	last 30
ACC	94.68	94.09	92.84	91.89	96.29	97.91	**98.42**	97.87	98.27
MCA	94.01	93.84	91.88	91.29	96.28	97.84	**98.57**	97.79	98.31

Table 6: The comparison with other approaches on the same dataset and with the same division of publicly available part of HEp-2 images. Presented values are in %.

	ACC	MCA
Gao et al. [10]	97.24	96.76
Shen et al. [28]	**98.82**	98.62
Our top performing method	98.53	**98.64**

Acknowledgement This research was supported by a research grant from the University of Southern Denmark, Odense University Hospital, Danish Cancer Society, and Region of Southern Denmark through the Project EFFICACY.

References

1. Babyak, M.: What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosomatic Medicine **66**(3), 411–421 (2004)
2. Baldi, P.: Deep Learning in Biomedical Data Science. Annual Review of Biomedical Data Science **1**, 181–205 (2018)
3. Bayramoglu, N., Kannala, J., Heikkilä, J.: Human Epithelial Type 2 Cell Classification with Convolutional Neural Networks. In: 15th Int. Conf. on Bioinformatics and Bioengineering. pp. 1–6. IEEE (2015)

4. Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z.: Deep Learning and its Applications in Biomedicine. Genomics, Proteomics & Bioinformatics **16**(1), 17–32 (2018)
5. Cascio, D., Taormina, V., Cipolla, M., Bruno, S., Fauci, F., Raso, G.: A Multi-Process System for HEp-2 Cells Classification Based on SVM. Pattern Recognition Letters **82**, 56–63 (2016)
6. Daněk, O., Matula, P., Maška, M., Kozubek, M.: Smooth Chan–Vese Segmentation via Graph Cuts. Pattern Recognition Letters **33**(10), 1405–1410 (2012)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE Conf. on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
8. Ensafi, S., Lu, S., Kassim, A.A., Tan, C.: Accurate HEp-2 Cell Classification Based on Sparse Coding of Superpixels. Pattern Recognition Letters **82**, 64–71 (2016)
9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A Guide to Deep Learning in Healthcare. Nature Medicine **25**(1), 24–29 (2019)
10. Gao, Z., Wang, L., Zhou, L., Zhang, J.: HEp-2 Cell Image Classification with Deep Convolutional Neural Networks. IEEE Journal of Biomedical and Health Informatics **21**(2), 416–428 (2017)
11. Harandi, M., Lovell, B., Percannella, G., Saggese, A., Vento, M., Wiliem, A.: Executable Thematic Special Issue on Pattern Recognition Techniques for Indirect Immunofluorescence Images Analysis. Pattern Recognition Letters **82**, 1–2 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
13. Hobson, P., Lovell, B., Percannella, G., Vento, M., Wiliem, A.: Benchmarking Human Epithelial Type 2 Interphase Cells Classification Methods on a Very Large Dataset. Artificial Intelligence in Medicine **65**(3), 239–250 (2015)
14. Hwang, S., Celebi, M.E.: Polyp Detection in Wireless Capsule Endoscopy Videos Based on Image Segmentation and Geometric Feature. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 678–681. IEEE (2010)
15. Iakovidis, D.K., Koulaouzidis, A.: Software for Enhanced Video Capsule Endoscopy: Challenges for Essential Progress. Nature Reviews Gastroenterology & Hepatology **12**(3), 172 (2015)
16. Kawahara, J., BenTaieb, A., Hamarneh, G.: Deep Features to Classify Skin Lesions. In: 13th Int. Symp. on Biomedical Imaging (ISBI). pp. 1397–1400. IEEE (2016)
17. Keogh, E., Mueen, A.: Curse of Dimensionality. Encyclopedia of Machine Learning and Data Mining pp. 314–315 (2017)
18. Ker, J., Wang, L., Rao, J., Lim, T.: Deep Learning Applications in Medical Image Analysis. IEEE Access **6**, 9375–9389 (2018)
19. Lai, Z., Deng, H.: Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron. Computational intelligence and neuroscience **2018** (2018)
20. Liu, A., Gao, Z., Tong, H., Su, Y., Yang, Z.: Sparse Coding Induced Transfer Learning for HEp-2 Cell Classification. Bio-medical Materials and Engineering **24**(1), 237–243 (2014)
21. Majtner, T., Yildirim-Yayilgan, S., Hardeberg, J.Y.: Combining Deep Learning and Hand-Crafted Features for Skin Lesion Classification. In: 6th Int. Conf. on Image Processing Theory, Tools and Applications. pp. 1–6. IEEE (2016)
22. Majtner, T., Yildirim-Yayilgan, S., Hardeberg, J.Y.: Optimised Deep Learning Features for Improved Melanoma Detection. Multimedia Tools and Applications **78**(9), 11883–11903 (2019)

23. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated Polyp Detection in Colon Capsule Endoscopy. IEEE Transactions on Medical Imaging **33**(7), 1488–1502 (2014)
24. Prasath, V.: Polyp Detection and Segmentation from Video Capsule Endoscopy: A review. Journal of Imaging **3**(1), 1–15 (2017)
25. Qi, X., Zhao, G., Chen, J., Pietikäinen, M.: HEp-2 Cell Classification: The Role of Gaussian Scale Space Theory as a Pre-processing Approach. Pattern Recognition Letters **82**, 36–43 (2016)
26. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434 (2015)
27. Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., Vaidya, V.: Understanding the Mechanisms of Deep Transfer Learning for Medical Images. In: Deep Learning and Data Labeling for Medical Applications, pp. 188–196. Springer (2016)
28. Shen, L., Jia, X., Li, Y.: Deep Cross Residual Network for HEp-2 Cell Staining Pattern Classification. Pattern Recognition **82**, 68–78 (2018)
29. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging **35**(5), 1285–1298 (2016)
30. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
32. Tajbakhsh, N., Shin, J., Gurudu, S., Hurst, R., Kendall, C., Gotway, M., Liang, J.: Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging **35**(5), 1299–1312 (2016)
33. Wainberg, M., Merico, D., Delong, A., Frey, B.J.: Deep Learning in Biomedicine. Nature Biotechnology **36**(9), 829 (2018)
34. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 399–407. Springer (2017)
35. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How Transferable are Features in Deep Neural Networks? In: Advances in Neural Information Processing Systems. pp. 3320–3328 (2014)
36. Yu, Y., Lin, H., Meng, J., Wei, X., Guo, H., Zhao, Z.: Deep Transfer Learning for Modality Classification of Medical Images. Information **8**(3), 91 (2017)
37. Yuan, Y., Li, B., Meng, M.Q.H.: Improved Bag of Feature for Automatic Polyp Detection in Wireless Capsule Endoscopy Images. IEEE Transactions on Automation Science and Engineering **13**(2), 529–535 (2016)