Deep Features for Breast Cancer Histopathological Image Classification

Fabio A. Spanhol [*], Paulo R. Cavalin [†], Luiz S. Oliveira [*], Caroline Petitjean [‡], and Laurent Heutte [‡]
[*] DInf, Federal University of Parana (UFPR), Curitiba, Brazil
Email: {faspanhol, lesoliveira}@inf.ufpr.br
[†] IBM Research, Rio de Janeiro, Brazil
Email: pcavalin@br.ibm.com
[‡] Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
Email: {caroline.petitjean, laurent.heutte}@univ-rouen.fr

Abstract—**Breast cancer (BC) is a deadly disease, killing millions of people every year. Developing automated malignant BC detection system applied on patient's imagery can help dealing with this problem more efficiently, making diagnosis more scalable and less prone to errors. Not less importantly, such kind of research can be extended to other types of cancer, making even more impact to help saving lives. Recent results on BC recognition show that Convolution Neural Networks (CNN) can achieve higher recognition rates than hand-crafted feature descriptors, but the price to pay is an increase in complexity to develop the system, requiring longer training time and specific expertise to fine-tune the architecture of the CNN. DeCAF (or deep) features consist of an in-between solution it is based on reusing a previously trained CNN only as feature vectors, which is then used as input for a classifier trained only for the new classification task. In the light of this, we present an evaluation of DeCaf features for BC recognition, in order to better understand how they compare to the other approaches. The experimental evaluation shows that these features can be a viable alternative to fast development of high-accuracy BC recognition systems, generally achieving better results than traditional hand-crafted textural descriptors and outperforming task-specific CNNs in some cases.**

I. Introduction

Cancer is currently a deadly disease rising across the globe. Some publications, such as that of the International Agency for Research on Cancer (IARC), which is part of the World Health Organization (WHO), report numbers of about 8.2 million deaths caused by cancer in the year of 2012 only. The incidence of this illness is expected to be of about 27 million new cases until 2030 [1]. Among the several existing types of cancer, breast cancer (BC) presents two very concerning characteristics: 1) it is the most common cancer among women worldwide; and 2) it presents a very high mortality rate when compared to other types of cancer. Since histopathological analysis remains the most widely used method for BC diagnosis [2], and most of the diagnosis continues being done by pathologists applying visual inspection of histological samples under the microscope, automatic classification of histopathological images is a research topic that can make BC diagnosis faster and less prone to errors. Until recently, though, works on BC histopathology image recognition systems have mainly worked with small datasets, which is generally a great limitation in developing high-accuracy image recognition systems. The recent release of the BreaKHis dataset [3], containing more than 7,900 images with four different magnifications from more than 80 patients, consisted of an important advance to bridge this gap, allowing researchers to apply the machine learning techniques for this problem.

Current state-of-art results on the BC recognition follow the two most common ways for designing image recognition systems. The approach in [3], which we generally refer to as *visual feature descriptors* or *hand-crafted features*, follows a more "traditional" approach, where an evaluation of the combination of six different feature sets and four base classifiers is conducted, and the final system is defined by the combination that produces the best results in the validation set. In contrast, in [4] and [5], the approaches follow the deep learning trend, where a Convolutional Neural Network (CNN) is trained for the BC recognition problem. The first is a method independent of magnification, based on single and multi-task CNN architectures. The second, referred here as either *CNN from scratch* or *task-specific CNN*, interchangeably, relies on the extraction of several small patches of the original images for training a specific CNN architecture. The reported results clearly show that the latter can achieve higher recognition rates. However, the development of such system requires longer training time, some tricks like random patches [6] to improve performance, and still a lot of expertise from the developer to tweak the system.

An in-between alternative to hand-crafted and task-specific CNN methods has been appearing frequently in the literature, often referred to as DeCAF features or neural codes [7]–[10]. This approach consists of reusing a pre-trained CNN only as a feature extractor, on top of which the parameters of a new classifier can be learned only for the new classification task. This approach has shown to be a very good general-purpose image feature extraction, providing competitive results in various tasks [7]–[10]. Although training a CNN from scratch when a large training set is available can still be the best option for the best accuracy, provided proper resources are available, DeCAF features can be a viable alternative to develop high-accuracy systems very fast, similar to a system based on hand-crafted features. Thus, if DeCAF features are able to outperform other visual feature descriptors, it can be set as a standard starting point to develop high-accuracy image recognition systems. And the development of accurate systems related to this area, e.g. systems for recognizing other types of cancer, can be done much faster.

Given these standpoints, the main focus of this work lies in evaluating DeCAF features for BC histopathological image classification, considering the BreaKHis dataset as benchmark, aiming at better understanding how this approach compares with hand-crafted descriptors and task-specific CNNs. More precisely, our goal is to make use of a pre-trained CNN to extract DeCAF features, from different layers of the network, to understand whether these features are good enough to compete with visual feature descriptors, such as those presented in [3], and how they compare with deep learning based methods, such as CNN trained from scratch for the problem, as in [4], and an independent magnification CNN approach, presented in [5]. To achieve these goals, we make use of the multiple feature vector (MFV) framework originally described in [11], which allows us also to evaluate this feature set in different scenarios, such as by combining classification results from sub-images (which we also refer to as patches) and/or from combining different feature sets. In this case, not only can we evaluate the performance of DeCAF features when a patch-based method is used, but also combine DeCAF features from different layers of the pre-trained CNN.

II. Related Work

In the literature, the first published work on automatic imaging processing for cancer diagnosis is dated more than 40 years [12]. Despite this long interest in this problem, developing solutions for it is still challenging due to the complexity of the images that such systems need to analyze.

The interest of the research community in this topic is proved by the high number of research papers published in recent years, related to this topic [13]–[16]. It is worth mentioning that most of these recent works related to BC classification are focused on Whole-Slide Imaging (WSI) [15]–[18]. However, the broad adoption of WSI and other forms of digital pathology has been facing obstacles such as the high cost of implementing and operating the technology, insufficient productivity for high-volume clinical routines, intrinsic technology-related concerns, unsolved regulatory issues, as well as “cultural resistance” from the pathologists [19].

Another relevant aspect is that until recently most of the works on BC histopathology image analysis were carried out on small datasets. Another drawback is that these datasets are usually not available to the scientific community, which not only makes it difficult for other researchers to develop new systems, since they need to gather images to compose the training set, but also to benchmark the results achieved by the systems. With the aim at bridging this gap, the BreaKHis dataset has been released and made freely available to the research community [3]. This database contains microscopic images from the surgical biopsy (SOB) of breast tumors, totalizing 7,909 images divided into benign and malignant tumors, which have been collected at four different magnification factors (or zoom level, which is a term that we make use of interchangeably): 40×, 100×, 200× and 400×. Samples have been generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE). These samples were prepared for histological study and labeled by pathologists of the Prevenção&Diagnose (P&D) Lab. The acquired digital images are available in 3-channel RGB (Red-Green-Blue) TrueColor (24-bit color depth, 8 bits per color channel) color space, dimension of 700 × 460 pixels. Figure 1 presents samples from this set, at the four corresponding magnification factors. A complete description of the BreaKHis database can be found in [3].

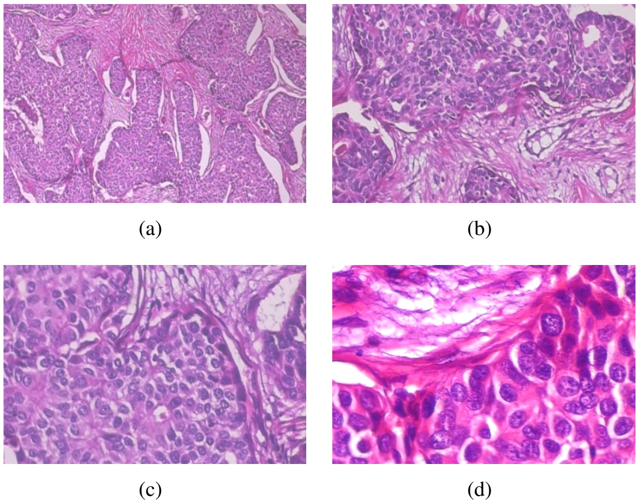
(a) (b) (c) (d)

Figure 1. Image samples from the BreaKHis database. Distinct areas, belonging to the same slide of breast malignant tumor (stained with HE), seen in different magnification factors: (a) 40×, (b) 100×, (c) 200×, and (d) 400×.

Since the recent publication of the BreaKHis dataset, some methods have been proposed using this dataset. In [3], the authors present an evaluation of different combinations of six different visual feature descriptors along with different classifiers. They report accuracies ranging from 80% to 85%, which may vary depending on the image magnification factor. Spanhol *et al.* [4] present results from a CNN for this set. Given that CNNs generally require large datasets, they make use of the random-patches trick, which consists of extracting sub-images at both training and test phases. During training, the idea is to increase the training set by means of extracting patches at randomly-defined positions. And during test, patches are extracted from a grid, and after classifying each patch, their classification results are combined. The authors show that, with this approach, increases in about 4 to 6 percentage points can be observed in the accuracy. Recently, Bayramoglu *et al.* [5] proposed a method to classify the BC histopathology images, which is independent of the magnifications factors. Their experimental results are competitive with previous state-of-the-art results obtained from hand-crafted features [3].

It is worth mentioning that deep learning approaches have been consistently outperforming more traditional machine learning methods in several tasks. Nonetheless, achieving good performance depends on the size of the training set, or on more specialized training schemes such as random patches, which generally require a very long training time. A solution that avoids having to handle large training datasets and long training time, and which has recently been reported with very good performance, is to rely on reusing existent pre-trained CNNs. Often referred to as DeCAF features or neural codes, this approach has been previously applied to diverse tasks, such as object recognition [7], image retrieval [8], texture recognition [9], among others [10].

III. DeCAF Features

The idea of DeCAF features consists of extracting features from an image and using them as input for a classifier, as any other feature set. Nevertheless, DeCAF are based on representation learning, where the parameters of a neural network are learned in a way that raw data, i.e., the pixels of the images, can be converted to a high-level representation [20]. The main difference between DeCAF features and the current standard of using CNNs [4], [6], [21], is that a previously-trained CNN is simply reused as feature extractor, the output of which is fed into another classifier, trained on problem-specific data.

In details, the DeCAF feature set consists of reusing the architecture and parameters of a pre-trained neural network, commonly a CNN, passing the input image through a feed-forward step, and using the outputs of a given layer of the network as input for the classifier [7]–[10]. To implement this idea, we make use of the pre-trained BVLC CaffeNet Model[1] (or CaffeNet for sake of simplicity), which is freely available on the Caffe deep learning framework[2]. This model consists of a slight modification of the AlexNet model [21], given that it has not been trained with data augmentation, and the order of the pooling and normalization layers is switched, i.e., in CaffeNet pooling is done prior to normalization.

The CaffeNet model has been trained on the ImageNet dataset [22], more specifically the dataset released for the ILSVRC12 challenge, obtaining a top-1 accuracy of 57.4% and a top-5 accuracy of 80.4% on the validation set. That set contains about 1.2 million samples, distributed into 1,000 distinct classes. Given the high number and variability of the classes, together with the high number of samples, the main assumption is that the representation learned from a CNN trained on this dataset defines a very good general-purpose feature extractor.

In order to convert the CaffeNet model into a feature extractor, we make use of the outputs of the top-most layers of the CNN, such as layers fc6, fc7, and fc8 (references are presented at the bottom right of Figure 2). The vectors corresponding to the output of those layers can then be used as inputs for a classifier, trained only on task-specific data.

IV. Experiments

In this section, we present an extensive experimental evaluation on the BreaKHis dataset, in order to evaluate DeCAF features in different scenarios. The accuracy is evaluated on each level of zoom independently, considering both the image-level and patient-level accuracy metrics. The reason for the second metric is that, generally, in medical imaging, decision is made patient-wise. For a better understanding, we define both metrics below.

Image-level accuracy simply corresponds to the score from the total number of correctly-classified images. That is, let N_{im} be the total number of images in the dataset, and N_c the total of correctly-classified images, image-level accuracy is defined as:

$$\text{Image-level accuracy} = \frac{N_c}{N_{im}}. \tag{1}$$

Patient-level accuracy, on the other hand, corresponds to the average image-level accuracy per patient. More formally, let N_P be the total number of patients, N_c^p be the total of correctly-classified images from patient p, and N_{im}^p the total of images for the same patient, patient-level accuracy is defined as:

$$\text{Patient-level accuracy} = \frac{\sum_{p=1}^{N_p} \frac{N_c^p}{N_{im}^p}}{N_P}. \tag{2}$$

Despite the relatively large number of layers in the CaffeNet model, in this work we focus only on extracting features from the three top-most layers, i.e., fc6, fc7, and fc8, which supposedly present the three most high-level features. These layers are composed of 4,096, 4,096, and 1,000 dimensions, respectively. Given the high-dimensionality of those vectors, we consider only Logistic Regression as base classifier, since it is fast at both training and classification phases, and can provide output probabilities.

The experiments have been organized in the following way. By considering a patch-based recognition and different configurations for it, we first evaluate the use of DeCAF feature sets individually, by making use of the output of either layer fc6, fc7, or fc8, and we consider systems with 1, 4 and 16 patches, based on the MFV framework presented in [11]. The main objective of these experiments is to observe the differences in accuracy of DeCAF features from different layers, and the impact when a patch-based classification is conducted.

Then, we conduct similar experiments, but considering the combination of more than one feature set at the same time, i.e. features from more than one layer of the network. Again, we implemented this idea based on the framework presented in [11], where the features are combined considering outputs at patch level.

For a direct comparison with the state of the art, the same partitions for the five-fold replications used in [3], [4] and available in the download page of the dataset[3].

[1] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

[2] http://caffe.berkeleyvision.org/

[3] http://web.inf.ufpr.br/vri/breast-cancer-database

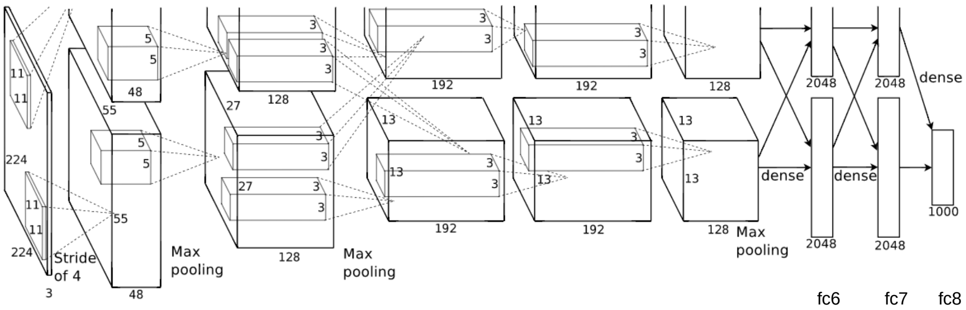

Figure 2. An illustration of the AlexNet model (extracted from [21]), which is used as baseline for the Caffe model. At the bottom right, the reference names for the top layers are listed.

Table I
ACCURACY, WITH RESPECTIVE STANDARD DEVIATION, WITHOUT COMBINATION OF LAYERS. **P** STANDS FOR PATIENT-LEVEL ACCURACY, **I** FOR IMAGE-LEVEL ACCURACY, AND #P FOR NUMBER OF PATCHES. IN BOLD, WITH A GRAY BACKGROUND, ARE HIGHLIGHTED THE BEST RESULTS AT EACH LEVEL AND MAGNIFICATION FACTOR.

	Layer	#p	Magnification factor			
			40×	100×	200×	400×
P	**fc8**	1	82.0 ± 2.5	82.0 ± 3.6	82.3 ± 2.1	81.3 ± 1.7
		4	82.3 ± 5.5	83.2 ± 6.2	81.6 ± 3.0	79.4 ± 6.3
		16	83.4 ± 6.9	83.8 ± 8.5	85.8 ± 3.5	80.7 ± 9.1
	fc7	1	83.1 ± 2.3	82.6 ± 3.5	82.5 ± 2.3	**81.9 ± 2.1**
		4	82.7 ± 5.0	83.0 ± 5.9	82.0 ± 2.8	80.4 ± 5.6
		16	**83.4 ± 6.7**	83.1 ± 8.4	**86.0 ± 3.7**	81.6 ± 8.6
	fc6	1	82.0 ± 3.3	83.3 ± 4.0	82.4 ± 3.1	81.0 ± 2.5
		4	82.8 ± 5.8	**83.9 ± 5.9**	81.8 ± 3.8	79.9 ± 6.1
		16	82.5 ± 8.6	83.6 ± 8.5	85.4 ± 5.2	81.1 ± 9.0
I	**fc8**	1	81.0 ± 1.6	80.9 ± 3.9	81.9 ± 1.1	80.2 ± 1.3
		4	83.7 ± 2.8	84.4 ± 4.3	82.0 ± 1.1	81.0 ± 2.6
		16	83.2 ± 2.4	84.0 ± 4.9	83.4 ± 1.1	80.9 ± 3.7
	fc7	1	82.2 ± 1.4	81.4 ± 3.9	81.9 ± 1.1	80.8 ± 1.5
		4	83.7 ± 2.7	83.7 ± 4.3	82.0 ± 1.1	81.4 ± 2.0
		16	83.1 ± 2.1	83.3 ± 4.6	**84.1 ± 1.5**	**81.6 ± 3.7**
	fc6	1	81.1 ± 2.3	82.1 ± 4.0	81.9 ± 1.4	79.8 ± 1.5
		4	**84.3 ± 2.9**	**84.7 ± 4.4**	82.2 ± 2.0	81.1 ± 2.2
		16	83.0 ± 2.6	84.6 ± 5.0	84.0 ± 2.8	81.1 ± 3.9

A. Results

By considering the previously-described setup, the first evaluation employed DeCAF features from each of the three aforementioned layers, individually, with 1, 4 and 16 patches. The results are listed in Table I. It is clear that features from layer fc8 perform worse than those from the other two layers, which presented the best results in all cases. Comparing to fc7 and fc6, there is a slight advantage for the first, with the best patient level accuracy in 3 out of 4 zoom levels, considering that both achieved the best accuracies at image level in two magnification factors. Regarding the use of patches, the results show that this could be an interesting alternative to improve the results with these features. Except from the $400\times$ zoom level, where the best patient-level accuracy was achieved with the entire image (a single patch), the best results in all the other zoom levels are with at least 4 patches. With zoom level $200\times$, the system with 16 patches performs considerably better.

B. Results Using Combination

The results presented herein are related to experiments evaluating the combination of DeCAF features from layers fc6, fc7, and fc8 (pointed out as simply 6, 7, and 8 given the space constraints), considering the four possible feature sets that could be used, i.e. 6+7+8, 6+7, 6+8, and, 7+8. Given that we have observed that combining features from the three layers simultaneously had not provided the highest recognition rates, Table II presents the results from only the pairwise combinations.

Overall, even though we can observe some improvements in the accuracy for some cases, the largest margin of gain, compared to the best results obtained with a single feature set at time, is of only 0.3%, i.e., the increase from 86.0% to 86.3% in patient accuracy in the $200\times$ magnification factor, and the increase from 84.3% to 84.6% in image accuracy in the $40\times$ magnification factor.

C. Comparison of Method Accuracy

In Table III we compare the accuracy of the approaches based on traditional hand-crafted features [3], task-specific CNN [4] and DeCAF features (this work). Such methods are compared in terms of F1 score (also referred to as F-score and F-measure in the literature [23]), which is given by the harmonic mean between precision and recall (Eq. 3),

$$\mathrm{F1} = 2 \times \frac{precision \times recall}{precision + recall}. \tag{3}$$

For patient-level evaluation, we consider the mean F1 score over all patients, similarly to patient-level accuracy defined in Equation 2. This metric can provide a better idea of the accuracy in detecting positive cases, i.e., malignant cancer, where errors in such detection are very costly for this sort of problem (it can cost patients' lives). In general, F1 score highlights better the nice performance of DeCAF features. Compared with the performance of the visual feature ex-

Table II
ACCURACY, WITH RESPECTIVE STANDARD DEVIATION, RESULTING FROM COMBINING DECAF FEATURES FROM LAYERS FC6, FC7 AND FC8. **P** STANDS FOR PATIENT-LEVEL ACCURACY, **I** FOR IMAGE-LEVEL ACCURACY, AND #P FOR NUMBER OF PATCHES. IN BOLD, WITH A GRAY BACKGROUND, ARE HIGHLIGHTED THE BEST RESULTS AT EACH LEVEL AND MAGNIFICATION FACTOR.

	Setup	#p	Magnification factor			
			40×	100×	200×	400×
P	**6+7**	1	82.6 ± 2.7	83.4 ± 4.4	82.7 ± 2.3	**82.1 ± 2.4**
		4	**83.6 ± 4.9**	83.4 ± 6.6	82.1 ± 3.4	80.6 ± 5.7
		16	83.4 ± 7.8	83.6 ± 8.7	85.8 ± 4.2	81.2 ± 9.0
	6+8	1	82.8 ± 2.6	83.5 ± 4.2	83.0 ± 2.1	82.0 ± 2.0
		4	83.4 ± 5.3	83.4 ± 6.6	82.2 ± 3.6	79.9 ± 6.1
		16	83.3 ± 7.4	**83.8 ± 8.7**	86.0 ± 4.1	80.7 ± 9.2
	7+8	1	82.7 ± 2.3	82.4 ± 3.4	82.8 ± 2.1	82.0 ± 1.5
		4	82.7 ± 5.0	83.3 ± 6.7	81.5 ± 3.2	79.9 ± 6.2
		16	83.3 ± 6.8	83.6 ± 8.6	**86.3 ± 3.5**	80.8 ± 9.0
I	**6+7**	1	81.7 ± 1.9	82.4 ± 4.5	82.2 ± 1.0	80.9 ± 1.6
		4	**84.6 ± 2.6**	84.4 ± 4.7	82.3 ± 1.9	**81.5 ± 2.6**
		16	83.4 ± 2.3	84.3 ± 5.0	84.1 ± 2.2	81.3 ± 3.9
	6+8	1	81.8 ± 1.8	82.4 ± 4.5	82.6 ± 1.0	80.8 ± 1.3
		4	84.5 ± 3.1	**84.8 ± 4.5**	82.6 ± 1.7	81.1 ± 2.5
		16	83.2 ± 2.4	84.8 ± 5.1	**84.1 ± 2.0**	81.0 ± 4.0
	7+8	1	81.8 ± 1.5	81.6 ± 4.0	82.2 ± 1.1	80.9 ± 1.0
		4	83.6 ± 2.8	84.4 ± 4.4	81.7 ± 1.3	81.4 ± 2.4
		16	83.0 ± 2.1	83.9 ± 4.7	84.0 ± 1.3	81.1 ± 3.8

tractors published in [3], our method outperforms the other approaches, at both patient and image level scores. Compared with the task-specific CNNs from [4], we can observe results that are similar to those with overall accuracy. However, a closer gap between the approaches is observed, especially at the 100× magnification factor.

Table III
F1 SCORES (PATIENT AND IMAGE LEVELS) OF THE THREE APPROACHES. THE BEST RESULTS ARE IN BOLD, AND ON A GRAY BACKGROUND ARE THE CASES AT WHICH HIGHER RESULTS ARE ACHIEVED IN THIS WORK, COMPARED THOSE PRESENTED IN [3]. RESULTS FROM THE COMBINATION OF MULTIPLE CLASSIFIERS ARE MARKED WITH *.

F1 score at	Approach	Magnification factor			
		40×	100×	200×	400×
Patient Level	[3]	86.0	84.9	87.8	85.6
	[4]	90.0	86.9	87.8	85.4
	[4]*	**93.5**	**91.7**	89.1	**89.9**
	This work	88.5	88.5	**90.3**	87.1
Image Level	[3]	87.8	86.1	88.5	86.3
	[4]	**92.9**	**88.9**	**88.7**	85.9
	[4]*	90.1	88.0	87.8	85.9
	This work	88.0	88.8	**88.7**	**86.7**

D. Discussion

For a better understanding of the results presented herein, in Table IV we compile the best results obtained in this work, and list them together with the best results presented in [3], [4] and [5]. All the results published in [5] are based on the patient score and the image level analysis is not available.

The main observation is that the use of DeCAF features can generally achieve better results than the use of more traditional visual feature descriptors, such as LBP (Local Binary Patterns) [24] and PFTAS (Parameter Free Threshold Analysis) [25], [26], and, in almost half of cases, even beat the results of a CNN [4], [5]. Compared with the traditional approach [3], only in the 200× zoom level there is a tie in image-level accuracy, while DeCAF loses in patient accuracy in the 400× magnification factor. In the remaining cases, the recognition rates achieved with DeCAF features are at least 0.4% better, but this difference can be as large as 4.1%. Compared with the CNN-based approach presented in [4], which achieved higher results, DeCAF features beat that method in the 200× zoom level, and in 400× in image accuracy. Without considering the combination of classifiers presented in [4], the system with DeCAF features also beats the CNN in patient accuracy in that magnification factor. And in the 40× magnification factor, image-level accuracy is close to that of CNN. However, for patient-level accuracy in the same zoom level, and both metrics in 100× magnification, CNN beats our results by a larger margin, ranging from 4.5 to 6.0%. This points out that the task-specific CNN might be better to deal with images with more fine-grained structures, while DeCAF features can be better suited for more coarse-grained problems.

Table IV
COMPARISON WITH THE LITERATURE. THE BEST RESULTS ARE IN BOLD, AND ON A GRAY BACKGROUND ARE THE CASES AT WHICH HIGHER RESULTS ARE ACHIEVED IN THIS WORK, COMPARED THOSE PRESENTED IN [3]. RESULTS FROM THE COMBINATION OF MULTIPLE CLASSIFIERS ARE MARKED WITH *.

%	Approach	Magnification factor			
		40×	100×	200×	400×
Patient	[3]	83.8 ± 4.1	82.1 ± 4.9	85.1 ± 3.1	82.3 ± 3.8
	[4]	88.6 ± 5.6	84.5 ± 2.4	85.3 ± 3.8	81.7 ± 4.9
	[5]	83.0 ± 3.0	83.1 ± 3.5	84.6 ± 2.7	82.1 ± 4.4
	[4]*	**90.0 ± 6.7**	**88.4 ± 4.8**	84.6 ± 4.2	**86.1 ± 6.2**
	This work	84.0 ± 6.9	83.9 ± 5.9	**86.3 ± 3.5**	82.1 ± 2.4
Image	[3]	82.8 ± 3.6	80.7 ± 4.9	**84.2 ± 1.6**	81.2 ± 3.6
	[4]	**89.6 ± 6.5**	**85.0 ± 4.8**	84.0 ± 3.2	80.8 ± 3.1
	[4]*	85.6 ± 4.8	83.5 ± 3.9	83.1 ± 1.9	80.8 ± 3.0
	This work	84.6 ± 2.9	84.8 ± 4.2	**84.2 ± 1.7**	**81.6 ± 3.7**

V. CONCLUSION

In this work we presented an investigation of the use of DeCAF features for breast cancer recognition using the BreaKHis dataset. The large size of the BreaKHis dataset has given us the opportunity to compare, on the same dataset, CNN trained from scratch with (DeCAF) features repurposed from another CNN trained on natural images, which often is not possible with medical image datasets since they are too small. From the results we can observe that these features are a viable alternative for a fast creation of image recognition systems using deep learning, and this system can perform better than systems using visual feature descriptors. Compared with a

CNN trained from scratch, DeCAF features present comparable recognition rates. Note that training a CNN specifically for the problem requires more complex and slower training schemes.

This result is important for the design of future classification based systems in computer-aided diagnosis, since it shows that deep learned features, even if obtained with a CNN trained on other types of images, are valuable. With this study we make one more step towards transfer learning for medical image analysis and CAD/CADx systems, as in [27], where CNN trained on ImageNet enable the detection of nodules in medical images.

As future work, one direction is to improve the recognition accuracy of DeCAF features using patches. Further investigation on the size of the patches, as well as overlapping patches, can be beneficial to increase the accuracies obtained with DeCAF features. Another investigation that can produce good results is the combination of these features with other visual descriptors and task-specific CNNs, to exploit the complementarity of these approaches. In addition, a better investigation on feature and classifier selection could also improve performance.

REFERENCES

[1] P. Boyle and B. Levin. (2008) World cancer report 2008. [Online]. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf

[2] S. R. Lakhani, E. I.O., S. Schnitt, P. Tan, and M. van de Vijver, *WHO classification of tumours of the breast*, 4th ed. Lyon: WHO Press, 2012.

[3] F. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering (TBME)*, vol. 63, pp. 1455–1462, 2015.

[4] ——, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2016, pp. 2560–2567.

[5] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *23rd International Conference on Pattern Recognition*, vol. 1, December 2016.

[6] L. G. Hafemann, L. E. S. Oliveira, and P. Cavalin, "Forest species recognition using deep convolutional neural networks," in *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, Aug. 2014, pp. 1103–1107.

[7] J. Donahue, J. Yangqing, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning (ICML)*. IMLS, Jun. 2014, pp. 647–655.

[8] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Cham: Springer International Publishing, 2014, ch. Neural Codes for Image Retrieval, pp. 584–599.

[9] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, 2016.

[10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, jun 2014, pp. 512–519.

[11] P. Cavalin, J. G. Martins, M. N. Kapp, and L. E. S. Oliveira, "A multiple feature vector framework for forest species recognition," in *Proceedings of the 28th Symposium on Applied Computing (SAC)*. ACM, Mar. 2013, pp. 16–20.

[12] B. Stenkvist, S. Westman-Naeser, J. Holmquist, B. Nordin, E. Bengtsson, J. Vegelius, O. Eriksson, and C. H. Fox, "Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," *Cancer Research*, vol. 38, no. 12, pp. 4688–4697, 1978.

[13] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Computers in Biology and Medicine*, vol. 43, no. 10, pp. 1563–1572, 2013.

[14] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2169–2178, 2013.

[15] Y. M. George, H. L. Zayed, M. I. Roushdy, and B. M. Elbagoury, "Remote computer-aided breast cancer detection and diagnosis system based on cytological images," *IEEE Systems Journal*, vol. 8, no. 3, pp. 949–964, 2014.

[16] Y. Zhang, B. Zhang, F. Coenen, J. Xiau, and W. Lu, "One-class kernel subspace ensemble for medical image classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 17, pp. 1–13, 2014.

[17] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1405–1420, 2013.

[18] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *in Proceedings of the 5th IEEE International Symposium on Biomedical Imageging (ISBI): From Nano to Macro*, vol. 61, May 2008, pp. 496–499.

[19] A. J. Evans, E. A. Krupinski, and L. Weinstein, Ronald S. Pantanowitz, "2014 american telemedicine association clinical guidelines for telepathology: Another important step in support of increased adoption of telepathology for patient care," *Journal of Pathology Informatics*, vol. 6, p. I13, 2015.

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

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2009, pp. 248–255.

[23] M. Sokolova, N. Japkowicz, and S. Szpakowicz, *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, pp. 1015–1021.

[24] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 24, pp. 971–987, 2002.

[25] N. A. Hamilton, R. S. Pantelic, K. Hanson, and R. D. Teasdale, "Fast automated cell phenotype image classification," *BMC Bioinformatics*, vol. 8, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/110

[26] L. P. Coelho, A. Ahmed, A. Arnold, J. Kangas, A. S.Sheikh, E. P. Xing, W. Cohen, and R. F. Murphy, "Structured literature image finder: extracting information from text and images in biomedical literature," in *Linking Literature, Information, and Knowledge for Biology*, ser. LNCS, C. Blaschke and H. Shatkay, Eds., 2010, vol. 6004, pp. 23–32.

[27] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. J. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *CoRR*, vol. abs/1602.03409, 2016. [Online]. Available: http://arxiv.org/abs/1602.03409