A Dataset for Breast Cancer Histopathological Image Classification

Fabio A. Spanhol*, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte

***Abstract*—Today, medical image analysis papers require solid experiments to prove the usefulness of proposed methods. However, experiments are often performed on data selected by the researchers, which may come from different institutions, scanners, and populations. Different evaluation measures may be used, making it difficult to compare the methods. In this paper, we introduce a dataset of 7909 breast cancer histopathology images acquired on 82 patients, which is now publicly available from http://web.inf.ufpr.br/vri/breast-cancer-database. The dataset includes both benign and malignant images. The task associated with this dataset is the automated classification of these images in two classes, which would be a valuable computer-aided diagnosis tool for the clinician. In order to assess the difficulty of this task, we show some preliminary results obtained with state-of-the-art image classification systems. The accuracy ranges from 80% to 85%, showing room for improvement is left. By providing this dataset and a standardized evaluation protocol to the scientific community, we hope to gather researchers in both the medical and the machine learning field to advance toward this clinical application.**

***Index Terms*—Breast cancer, histopathology, image classification, medical imaging.**

I. Introduction

CANCER is a significant public health problem in the world today. According to the International Agency for Research on Cancer of the World Health Organization, 8.2 million deaths were caused by cancer in 2012 and 27 million of new cases of this disease are expected before 2030 [1]. In particular, breast cancer (BC) is one of most common types of cancer among women. Mortality of BC is very high when compared to other types of cancer.

Detection and diagnosis of BC can be achieved by imaging procedures such as diagnostic mammograms (X-rays), magnetic resonance imaging, ultrasound (sonography), and thermography [2]. Imaging for cancer screening has been investigated for more than four decades [3]. However, biopsy is the only way to diagnose with confidence if cancer is really present. Among biopsy techniques, the most common are fine needle aspiration, core needle biopsy, vacuum-assisted, and surgical (open) biopsy (SOB) [4]. The procedure consists in collecting samples of cells or tissue, which are fixed across a glass microscope slide for subsequent staining and microscopic examination. Diagnosis from a histopathology image is thus the gold standard in diagnosing almost all types of cancer, including BC [5], [6]. The final BC diagnosis, including grading and staging, is done by pathologists applying visual inspection of histological samples under microscope.

Histopathological analysis is a highly time-consuming specialized task, dependent on the experience of the pathologists and influenced by factors such as fatigue and decrease of attention. As pointed by Gurcan *et al.* [7], there is a pressing need for computer-assisted diagnosis (CAD) to relieve the workload on pathologists by filtering obviously benign areas, so that the experts can focus on the more difficult-to-diagnose cases [8].

A considerable amount of efforts has thus been devoted to the field of BC histopathology image analysis, and in particular to the automated classification of benign or malignant images, for computer-aided diagnosis. Kowal *et al.* [9] compare and test different algorithms for nuclei segmentation on a dataset of 500 images, for which accuracies ranging from 96% to 100% are reported. Filipczuk *et al.* [10] present a BC diagnosis system based on the analysis of cytological images of fine needle biopsies, to discriminate the images as either benign or malignant. Using four different classifiers trained with a 25-D feature vector, they report a performance of 98% on 737 images. Similarly to [9] and [10], George *et al.* [11] propose a diagnosis system for BC based on the nuclei segmentation of cytological images. Using different machine learning models, such as neural networks and support vector machines (SVMs), they report accuracy rates ranging from 76% to 94% on a dataset of 92 images. Zhang *et al.* [12] propose a cascade approach with rejection option. In the first level of the cascade, authors expect to solve the easy cases, while the hard ones are sent to a second level where a more complex pattern classification system is used. They assess the proposed method on a database proposed by the Israel Institute of Technology, which is composed of 361 images (40× magnification). On this dataset, they report results of 97% of reliability. In another work [13], the same authors assessed an ensemble of one-class classifiers on the same database achieving a recognition rate of 92%.

We can gather from the literature that most of the works on BC histopathology image analysis are carried out on small datasets, which are usually not available to the scientific community. In a recent review, Veta *et al.* [14] point out that the main obstacle in the development of new histopathology image analysis methods is the lack of large, publicly available, annotated datasets. Annotated database is also crucial to develop and validate machine learning systems.

Manuscript received May 14, 2015; revised August 18, 2015 and October 21, 2015; accepted October 22, 2015. Date of publication October 30, 2015; date of current version June 16, 2016. This work was supported by the National Council for Scientific and Technological Development (CNPq), Brazil, under grant #301653/2011-9 and the Coordination for the Improvement of Higher Level Personnel (CAPES) through SticAmsud project #263/2014. *Asterisk indicates corresponding author.*

*F. A. Spanhol is with the Federal University of Parana, Curitiba-PR 80060-000, Brazil (e-mail: faspanhol@inf.ufpr.br).

L. S. Oliveira is with the Federal University of Parana.

C. Petitjean and L. Heutte are with LITIS EA 4108, Université de Rouen.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2015.2496264

In this paper, we introduce a database, called BreaKHis, that is intended to mitigate this gap. BreaKHis is composed of 7909 clinically representative microscopic images of breast tumor tissue images collected from 82 patients using different magnifying factors (40×, 100×, 200×, and 400×). To date, it contains 2480 benign and 5429 malignant samples. This database has been built in collaboration with the P&D Laboratory[1]—Pathological Anatomy and Cytopathology, Parana, Brazil. We believe that researchers will find this database useful as it makes future benchmarking and evaluation possible. The database is available for research purposes from now on, upon request.[2]

Additionally, we present in this paper the classification performance of a baseline pattern recognition system, designed to discriminate between benign and malignant tumors with state-of-the-art feature extractors and classifiers, with the aim of showing the difficulty of the problem. The classification system is based on four machine learning models, trained with different textural representations and keypoint detectors. A comprehensive set of experiments shows that accuracy rates with this baseline system range from 80% to 85%, depending on the image magnification factor. To give an insight about the discriminative power of the textural representations we have used, we also present the performance of the oracle. The oracle is an abstract model defined in [15], which always selects the classifier that predicted the correct label, for a given query sample, if such a classifier exists. In other words, it represents the ideal classifier selection scheme. The difference between the performance of a real-life classification system and the abstract model of the oracle shows that room for improvement is left with a high potential of increased accuracy. Performance may be improved by using dedicated, improved descriptors, or designing a strategy to select appropriate descriptors.

This paper is structured as follows. Section II introduces the proposed database. Section III describes the feature sets and the classifiers. Section IV reports our experiments and discusses our results. Finally, Section V concludes the work.

II. BreaKHis Dataset

The BreaKHis database contains microscopic biopsy images of benign and malignant breast tumors. Images were collected through a clinical study from January 2014 to December 2014. All patients referred to the P&D Laboratory, Brazil, during this period of time, with a clinical indication of BC were invited to participate in the study. The institutional review board approved the study and all patients gave written informed consent. All the data were anonymized.

Samples are generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE). The samples are collected by SOB, prepared for histological study, and labeled by pathologists of the P&D Lab. The preparation procedure used in this work is the standard paraffin process, which is widely used in clinical routine. The main goal is to preserve the original tissue structure and molecular composition, allowing to observe it in a light microscope. The complete preparation procedure includes steps such as fixation, dehydration, clearing, infiltration, embedding, and trimming [16]. To be mounted on slides, sections of 3 μm are cut using a microtome. After staining, the sections are covered with a glass coverslip. Then, the anatomopathologists identify the tumoral areas in each slide, by visual analysis of tissue sections under a microscope. Final diagnosis of each case is produced by experienced pathologists and confirmed by complementary exams such as immunohistochemistry analysis.

TABLE I
Magnification and Digital Resolution of the Acquisition System

Visual magnification	Objective lens	Effective pixel size (μm)
40×	4×	0.49
100×	10×	0.20
200×	20×	0.10
400×	40×	0.05

An Olympus BX-50 system microscope with a relay lens with magnification of 3.3× coupled to a Samsung digital color camera SCC-131AN is used to obtain digitized images from the breast tissue slides. This camera uses a 1/3" Sony Super-HAD (Hole-Accumulation Diode) interline transfer charge-coupled device with pixel size 6.5 μm × 6.25 μm and a total pixel number of 752 × 582. Images are acquired in three-channel red–green–blue (RGB) TrueColor (24-bit color depth, 8 bits per color channel) color space using magnifying factors of 40×, 100×, 200×, and 400×, corresponding to objective lens 4×, 10×, 20×, and 40×. The camera is set for automatic exposure and focusing is done manually on the microscope looking at the digital image on the computer screen. Table I shows the effective pixel size in micrometers for each magnifying factor and objective lens we have used. The pixel size is the physical pixel size of the camera (6.5 μm), divided by the relay lens magnification (3.3) and the objective lens.

The original images contain black borders on both the left and right sides and text annotations in the upper left corner. To remove these undesired areas, the resulting images are cropped and saved in three-channel RGB, 8-bit depth in each channel, portable network graphics format with no compression, dimension of 700 × 460 pixels. Resulting images are raw images without normalization nor color standardization.

The acquisition of images at different magnifications is performed as follows: first the pathologist identifies the tumor and defines a region of interest (ROI). To cover the whole ROI, several images are captured using the lowest magnification, i.e., 40×. The pathologist preferentially selects images with a single type of tumor (majority of the cases), but some of the images also include transitional tissue, e.g., normal-pathological. In average, a total of 24 images per patient is captured from each slide using the lowest magnification (see Table II). Then, the magnification is manually increased to 100× and a similar number of images is captured inside the initial ROI. This process is repeated for 200× and 400× magnifications, respectively. A final visual (i.e., manual) inspection discards out-of-focus images.

[1] http://www.prevencaoediagnose.com.br/

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

TABLE II
IMAGE DISTRIBUTION BY MAGNIFICATION FACTOR AND CLASS

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909
# Patients	24	58	82

TABLE III
BENIGN IMAGE DISTRIBUTION BY MAGNIFICATION FACTOR AND HISTOLOGICAL SUBTYPES

Magnification	A	F	TA	PT	Total
40×	114	253	109	149	598
100×	113	260	121	150	614
200×	111	264	108	140	594
400×	106	237	115	130	562
Total	444	1014	453	569	2368
# Patients	4	10	3	7	24

TABLE IV
MALIGNANT IMAGE DISTRIBUTION BY MAGNIFICATION FACTOR AND HISTOLOGICAL SUBTYPES

Magnification	DC	LC	MC	PC	Total
40×	864	156	205	145	1370
100×	903	170	222	142	1437
200×	896	163	196	135	1390
400×	788	137	169	138	1232
Total	3451	626	792	560	5429
# Patients	38	5	9	6	58

To date, the database is composed of 7909 images divided into benign and malignant tumors. Table II summarizes the image distribution. Both breast tumors, benign and malignant, can be sorted into different types based on the aspect of the tumoral cells under the microscope. The dataset currently contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant tumors (breast cancer): ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The distribution of benign and malignant tumors in these classes is presented in Tables III and IV, respectively.

Fig. 1 shows four images—with the four magnification factors (a) 40×, (b) 100×, (c) 200×, and (d) 400×—acquired from a single slide of breast tissue containing a malignant tumor (breast cancer). Highlighted rectangle (manually added for illustrative purposes only) is the area of interest selected by pathologist to be detailed in the next higher magnification.

III. FEATURE EXTRACTORS AND CLASSIFIERS

Histological tissue images can be characterized by two types of approaches. The first one is based on explicit segmentation to extract structure properties, such as nuclei shape, glandular unit shape, etc., while the second one is a global approach based on texture representation. Since segmentation of histological tissue images is not a trivial task and can be prone to errors, we have chosen a global approach based on state-of-the-art texture representation.

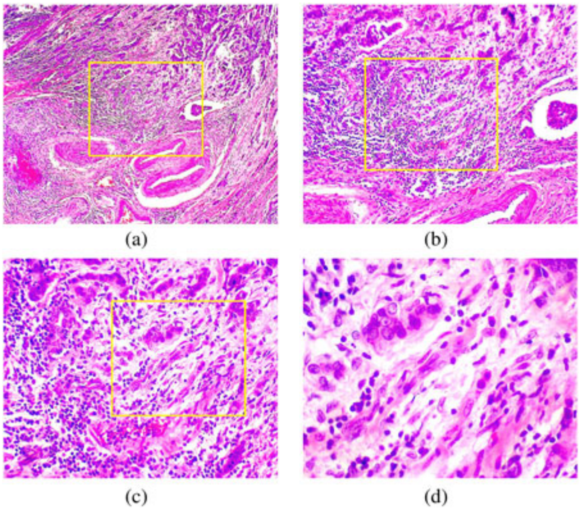

Fig. 1. Slide of breast malignant tumor (stained with HE) seen in different magnification factors: (a) 40×, (b) 100×, (c) 200×, and (d) 400×. Highlighted rectangle (manually added for illustrative purposes only) is the area of interest selected by pathologist to be detailed in the next higher magnification factor.

In this section, we briefly describe all the representations we have used to train the classifiers. These include the textural descriptors most commonly found in the literature, such as local binary patterns (LBP) [17], completed LBP (CLBP) [18], local phase quantization (LPQ) [19], gray-level co-occurrence matrix (GLCM) [20], threshold adjacency statistics (TAS) [21], and one keypoint descriptor, named ORB [22]. Keypoint descriptors are most often used for object recognition; however, the literature shows that this kind of descriptor can produce interesting results for texture classification on microscopic images [23].

A. Local Binary Patterns

The LBP operator [17] consists in computing the distribution of binary patterns in the circular neighborhood of each pixel. The neighborhood is characterized by a radius R and a number of neighbors P. The principle is to threshold neighboring pixels, compared to the central pixel: to each of the P neighbors, the value 1 is assigned, if the current pixel intensity is superior or equal to the central pixel intensity; otherwise, value 0 is assigned. Thus, for each pixel, a binary pattern is obtained from the neighborhood. A total of 2^P different binary patterns can be obtained. The LBP code at pixel p is obtained by computing the scalar product between the binary code and a vector of powers of two, and summing up the result:

$$\mathrm{LBP}(p) = \sum_{i=0}^{P-1} 2^i . \delta(f(q_i) - f(p)) \tag{1}$$

where $f(q_i)$ and $f(p)$ are gray levels of pixels q_i and p, respectively, and δ is the Kronecker function. Histogram of the LBP codes can then be used as a texture descriptor. Note that some patterns, which are identical up to one or several rotations, do not have the same LBP code: for example, 10000000 and 01000000 have 255 and 128 as LBP codes, respectively. This behavior can be avoided with the rotation invariant LBP, introduced in [17]: each pattern is rotated P times, and the minimum LBP code over the P rotations is retained. With this modification, 1000000 and 01000000 have the same LBP code. For $P = 8$, the rotation invariant LBP method decreases the number of possible patterns from 256 to 36.

Another improvement originated from the observation that some binary patterns occur more often in texture images than others. These frequent patterns are usually those with a small number of transitions, i.e., 0-1 or 1-0: for example, 00000000 (no transition), 011111111 (two transitions), 00011111 (two transitions) occur more often than 10101010 (eight transitions), or 01100101 (six transitions). The frequent patterns are called uniform patterns [17]. The LBP method that takes into account uniform patterns makes the number of LBP codes used for histogram bins decrease from 36 to 10. In our experiments, we work with rotation-invariant uniform patterns, with a standard value of $P = 8$ neighbors, providing a 10-D feature vector.

B. Completed Local Binary Pattern

The CLBP is one of the latest variants of LBP is the CLBP [18], which provides a completed modeling of the LBP, based on three components extracted from the local region: center pixel, sign, and magnitude. The center pixel is coded by a binary code after global thresholding, with the threshold set as the average gray level of the whole image. For the two other components, a neighborhood of radius R and number of neighbors P is considered, similarly to LBP. The difference signs and magnitudes are then computed and coded by specific operator into binary format so that they can be combined to form the final CLBP histogram [18]. Note that the operator coding the sign component corresponds to the original LBP operator. We have assessed different configurations suggested in [18] and the best results observed in our experiments have been obtained with the combination of all components using a 3-D joint histogram, while the best values for the parameters P and R are 24 and 5, respectively, yielding a 1352-D feature vector.

C. Local Phase Quantization

LPQ is based on quantized phase information of the discrete Fourier transform (DFT) [19]. It uses the local phase information extracted using the 2-D DFT or, more precisely, a short-term Fourier transform computed over a rectangular $M \times M$ neighborhood N_p at each pixel position p of the image $f(p)$. The quantized coefficients are represented as integer values in the range 0–255 using binary coding described in [19]. These binary codes are generated and accumulated in a 256-bin histogram, similar to the LBP method. The accumulated values in the histogram are used as the LPQ 256-D feature vector. In our experiments, a variant of LPQ, named LPQ-TOP [24], produced better results. The main difference is that LPQ and LPQ-TOP use different default values for their parameters.

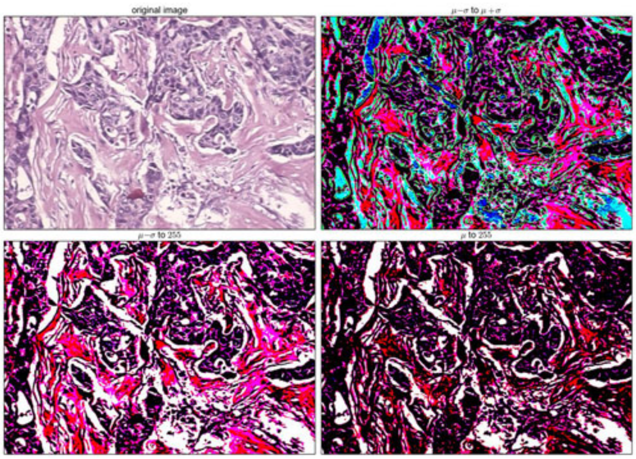

Fig. 2. PFTAS thresholding on a malignant image. From left to right, top to bottom: original image, binarized images using threshold ranges [$\mu + \sigma$, $\mu - \sigma$], [$\mu - \sigma$, 255], and [μ, 255].

D. Gray-Level Co-Occurrence Matrices

GLCM are widely used to characterize texture images. In our experiments, four adjacency directions $0°$, $45°$, $90°$, $135°$, and eight gray levels are used to compute the GLCM. On the GLCM, 13 Haralick parameters are computed [20]: angular second moment, contrast, correlation, sum of squares, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation 1, and information measures of correlation 2. Finally, we obtain a final feature vector by averaging the 13-D feature vectors in the four directions.

E. Parameter-Free Threshold Adjacency Statistics (PFTAS)

The TAS is a simple and fast morphological measure introduced in [21] for cell phenotype image classification. Since BC images share some similarities with these images, we have used the PFTAS [25], the parameter-free version of TAS. Its principle is to accumulate in the histogram bins, pixels according to their number of white neighbors, in multiple-threshold binarized images. The original image is binarized using three different threshold ranges: [$\mu + \sigma, \mu - \sigma$], [$\mu - \sigma, 255$], and [$\mu$, 255], where μ is an Otsu defined threshold, and σ is the standard deviation of the above threshold pixels. Fig. 2 illustrates these images.

For each binarized image, a normalized histogram of pixels having i (i ranging from 0 to 8) white pixels as neighbors is computed. All three histograms are concatenated to form a 27-D feature vector for each one of three RGB channels, yielding a 81-D feature vector. Finally, this vector and its bitwise negated version are concatenated, resulting in a 162-D feature vector.

F. ORB

ORB (for Oriented FAST and Rotated BRIEF) [22] has been proposed as an alternative to the traditional SIFT [26] and SURF [27] keypoint detectors, in terms of computational cost and matching performance. It is designed to be rotation invariant and resistant to noise. ORB is based on the well-known FAST keypoint detector [28] and the BRIEF keypoint descriptor [29]. ORB works as follows: first FAST is used to find keypoints; then, Harris corner detection selects the top N points among them. Since FAST features do not have an orientation component, an efficiently computed orientation is added. This orientation compensation mechanism makes ORB rotation invariant.

In this work, we have used the OpenCV implementation [30] with the default parameters, which returns a 32-D vector for each keypoint. Best results have been achieved using 500 keypoints, considering balance between runtime and improvement of recognition rate. At the end, the image is represented by a single 32-D vector that contains the average of all keypoints.

G. Classifiers

Four different classifiers were used to assess the aforementioned feature sets: a 1-nearest neighbor (1-NN), quadratic linear analysis (QDA), SVMs, and random forests (RF) of decision trees. A k-NN is a type of instance-based learning that stores all available training data and classifies the testing samples based on a similarity measure (e.g., Euclidean distance). In particular, the 1-NN is often used to assess the discriminating power of the features. QDA is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements from each class are normally distributed. Unlike LDA however, QDA does not assume that the covariance of each of the classes is identical. SVM, a very popular classification algorithm, builds a hyperplane in a high-dimensional space, which can be used for classification and regression. Differently from other linear discriminant functions, it provides the optimal hyperplane that separates two classes [31]. RF is an ensemble approach that combines decision tree predictors. The principle behind ensemble methods is that a group of weak learners (in this case the decision trees) can come together to form a strong learner [32]. One of the advantages of the RF is that they are quite fast and able to deal with unbalanced data.

IV. Experiments and Discussion

A. Protocol

The BreaKHis dataset has been randomly divided into a training (70%) and a testing (30%) set. To make sure the classifier generalizes to unseen patients, we guarantee that patients used to build the training set are not used for the testing set. The results presented in this work are the average of five trials. This protocol was applied independently to each of the four magnifications available. Note that we have also made the folds available along with the dataset, to allow for a full comparison of classification results.

For the SVM, different kernels were tried; we retained the Gaussian kernel which produced the best results. The kernel parameters γ and C were empirically defined through a grid search and fivefold cross-validation using the training set. The same protocol was applied to tune the parameters of the RF. All the experiments were carried out using scikit-learn, an open-source machine learning library in Python [33]. Table V recalls the six representations we have used to train the classifiers.

TABLE V
SUMMARY OF THE DESCRIPTORS

Name	Feature number
CLBP	1352
GLCM	13
LBP	10
LPQ	256
ORB	32
PFTAS	162

Since the decision is patientwise, we report the recognition rate at the patient level, and not at the image level. Let N_P be the cancer images of patient P. For each patient, if N_{rec} images are correctly classified, then one can define a patient score as

$$\text{Patient Score} = \frac{N_{\text{rec}}}{N_P} \tag{2}$$

and the global recognition rate as

$$\text{Recognition Rate} = \frac{\sum \text{Patient score}}{\text{Total number of patients}}. \tag{3}$$

The receiver operating characteristic (ROC) curve is another valuable tool for performance analysis, especially since our data are unbalanced data. Indeed, the ROC curve is insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change [34].

B. Results

Table VI reports the performance of all classifiers and descriptors we have assessed. We propose a two-level analysis of this table. Let us first focus on the influence of the magnification factors, by comparing columns (best results in bold). Interestingly, the magnification factors do not seem to have the same level of information. In particular, the first level (40×) exhibits the best results over CLBP, LBP, and ORB. This slight tendency that 40× may be the most informative magnification factor is in accordance with the pathologist behavior, which starts by examining factor 40 and switches to the next level, until he establishes his diagnosis. Note, however, that the 200× magnification factor also shows high potential, with the best results over GLCM and PFTAS, higher than those obtained with the 40× level. The complementarity of the magnification factors could be fruitfully investigated in the future, through a coarse-to-fine analysis for example. It is beyond the scope of this paper.

The other level of analysis concerns the feature vector comparison (best results are underlined in Table VI). All feature vectors exhibit stable and close results. These results are little influenced by the classifiers: for each factor and for each feature vector, the recognition rates of the four classifiers are in a range of less than 4%. Note, however, that the results obtained by CLBP with QDA fall out of this range and are far below the other mean recognition rates. Indeed, QDA is based on the estimation of covariance matrices: in order to make a proper estimation of these matrices, a large amount of samples is

TABLE VI
MEAN RECOGNITION RATES AND STANDARD DEVIATIONS OF THE CLASSIFIERS TRAINED WITH DIFFERENT DESCRIPTORS

Descriptor	Classifier	Magnification Factors			
		40×	100×	200×	400×
CLBP	1-NN	**73.6** ± 2.5	71.0 ± 2.8	69.4 ± 1.5	70.1 ± 1.3
	QDA	39.4 ± 13.5	**51.7** ± 17.3	50.3 ± 16.0	49.4 ± 15.5
	RF	**74.5** ± 0.7	72.5 ± 3.8	70.0 ± 2.4	72.3 ± 2.1
	SVM	**77.4** ± 3.8	76.4 ± 4.5	70.2 ± 3.6	72.8 ± 4.9
GLCM	1-NN	74.7 ± 1.0	76.8 ± 2.1	**83.4** ± 3.3	81.7 ± 3.3
	QDA	67.0 ± 6.0	74.2 ± 3.5	**78.6** ± 1.7	77.0 ± 2.3
	RF	73.6 ± 1.5	76.0 ± 1.9	**82.4** ± 2.3	79.8 ± 2.5
	SVM	74.0 ± 1.3	78.6 ± 2.6	**81.9** ± 4.9	81.1 ± 3.2
LBP	1-NN	**75.6** ± 2.4	73.0 ± 2.4	72.9 ± 2.3	71.2 ± 3.6
	QDA	69.7 ± 3.8	69.7 ± 4.2	68.8 ± 4.7	**72.3** ± 4.6
	RF	**74.0** ± 2.9	73.1 ± 1.9	70.1 ± 2.5	70.7 ± 4.3
	SVM	**74.2** ± 5.0	73.2 ± 3.5	71.3 ± 4.0	73.1 ± 5.7
LPQ	1-NN	72.8 ± 4.9	71.1 ± 6.4	**74.3** ± 6.3	71.4 ± 5.2
	QDA	**70.4** ± 1.1	69.3 ± 4.2	67.2 ± 1.9	68.3 ± 1.8
	RF	**73.8** ± 5.0	72.3 ± 5.5	73.4 ± 5.9	71.1 ± 3.8
	SVM	**73.7** ± 5.5	72.8 ± 5.0	73.0 ± 6.6	**73.7** ± 5.7
ORB	1-NN	**71.6** ± 2.0	69.3 ± 2.0	69.6 ± 3.0	66.1 ± 3.5
	QDA	**74.4** ± 1.7	66.5 ± 3.2	63.5 ± 2.7	63.5 ± 2.2
	RF	**72.3** ± 1.8	69.3 ± 1.0	68.6 ± 1.7	67.6 ± 1.2
	SVM	**71.9** ± 2.3	69.4 ± 0.4	68.7 ± 0.8	67.3 ± 3.1
PFTAS	1-NN	80.9 ± 2.0	80.7 ± 2.4	**81.5** ± 2.7	79.4 ± 3.9
	QDA	83.8 ± 4.1	82.1 ± 4.9	**84.2** ± 4.1	82.0 ± 5.9
	RF	81.8 ± 2.0	81.3 ± 2.8	**83.5** ± 2.3	81.0 ± 3.8
	SVM	81.6 ± 3.0	79.9 ± 5.4	**85.1** ± 3.1	82.3 ± 3.8

Bold shows the best results over the magnification factors. For each magnification factor, underlining shows the five best results over the feature vectors and classifiers.

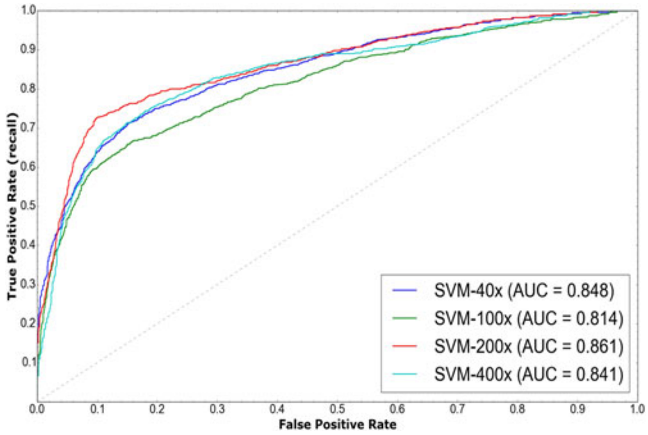

Fig. 3. ROC curves for the confusion matrices presented in Table VII.

TABLE VII
CONFUSION MATRICES PRODUCED BY THE SVM CLASSIFIER TRAINED WITH THE PFTAS DESCRIPTOR

	40×		100×		200×		400×	
	B	M	B	M	B	M	B	M
B	0.62	0.38	0.38	0.62	0.75	0.25	0.75	0.25
M	0.06	0.94	0.06	0.94	0.06	0.94	0.11	0.89

B: benign, M: malignant.

TABLE VIII
ERROR DISTRIBUTION (%) OF THE SVM TRAINED WITH PFTAS OVER SUBCLASSES

Class	Subclass	Magnification Factors			
		40×	100×	200×	400×
Benign	Adenosis	15.7	21.7	9.7	10.3
	Fibroadenoma	**28.5**	**31.8**	**29.5**	**30.2**
	Phyllodes Tumor	13.6	18.6	10.1	14.4
	Tubular Adenoma	23.1	19.5	15.6	16.5
Malignant	Ductal	11.6	2.8	13.9	8.7
	Lobular	0.0	0.0	0.2	3.2
	Mucinous	2.8	5.1	13.9	10.1
	Papillary	4.7	0.5	7.1	6.6

A large amount of false positive comes from fibroadenoma (benign) mistaken for malignant tumor.

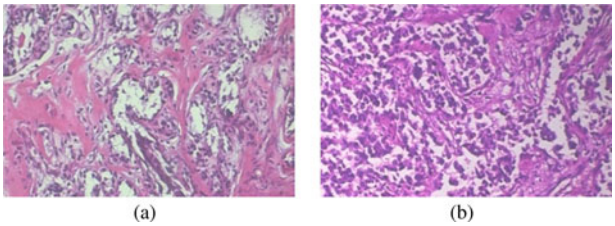

Fig. 4. Example of misclassification: (a) benign tumor classified as a malignant tumor and (b) real malignant tumor.

required, which should be all the greater given that CLBP is high dimensional (1352).

Over all the feature vectors, the PFTAS performs best. Since the best overall performance (recognition rate of 85.1% for factor 200×) is achieved by the SVM trained with PFTAS descriptors, we focus on the SVM/PFTAS association and further analyze their performance, by drawing the associated ROC curve (see Fig. 3) and reporting the confusion matrices in Table VII, which confirms that 200 seems to be the most discriminant magnification factor. As we can see, most of the confusions occur when a benign tumor is classified as malignant (high false positive rate). This may be partially explained, as pointed out by Kowal *et al.* [9], by the fact that one of the benign tumor present in the dataset (fibroadenoma) shares similar properties with a malignant tumor. To verify this hypothesis, we analyze the origin of errors in the SVM/PFTAS results in Table VIII. This analysis shows that independently of the magnification factor, about 30% of errors of the classifier are due to benign tumors fibroadenoma classified as malignant class. One example of this misclassification is presented in Fig. 4, where (a) shows a benign tumor classified as a malignant tumor and (b) presents a real malignant tumor.

In spite of the complexity of the problem, a reliable CAD system should produce very low false positive and negative rates. This will be the main challenge for researchers willing to use the proposed dataset. One way to build a more reliable system is by combining the classifiers into a multiple classifier system framework [35]. Another approach that has gained a lot of attention in the pattern recognition community recently is the dynamic selection of classifiers (DSC), which selects a different classifier for each new test sample. DSC techniques rely on the assumption that each base classifier is an expert

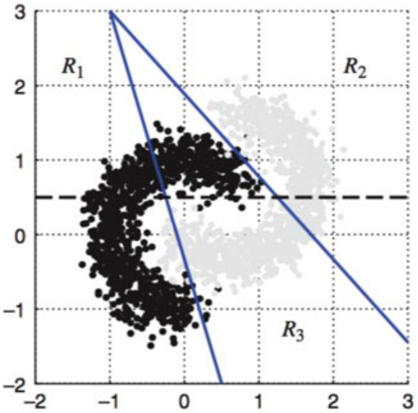

Fig. 5. Feature space partitioned into three competence regions [36]. Blue lines delimit local region in which a competent classifier can be found.

in a different local region of the feature space. Based on this hypothesis, these techniques try to select the most competent classifier for the local region in the feature space where the test sample is located.

To show why classifier selection works, we use the example presented by Kuncheva in [36]. Consider the two-class problem depicted in Fig. 5 and a pool D of three weak classifiers, $D = \{D_1, D_2, D_3\}$. Suppose that D_1 always predicts class "black" and that D_2 always predicts class "gray." D_3 is a linear classifier whose discriminant function is shown as the horizontal dashed line in Fig. 5. D_3 predicts class "black" for samples above the line and class "gray" for samples underneath. The individual accuracy of these classifiers is about 50%; therefore, the majority vote among them is useless as it will always match the decision of the arbiter D_3 and lead to 50% error. However, if we use the three local regions, delimited by the blue lines, and nominate the most competent classifier for each region (D_1 in R_1, D_2 in R_2, D_3 in R_3), the error of the ensemble will be negligible.

This example shows the potential of the DSC approach. In real life, it may be quite difficult to find regions that have such a huge impact on the ensemble performance [36]. The literature shows several different methods to define such regions. A recent review can be found in [37].

To assess the potential of the DSC approach, i.e., to verify a given pool of classifiers is competent, a common method is to compute the accuracy of the oracle, which is the upper limit in terms of performance of the pool of classifiers. As stated in Section I, the oracle is an abstract model which always selects the classifier that predicted the correct label, for a given query sample, if such a classifier exists.

A good oracle does not necessarily imply a good performance on a real-life classification system. However, a DSC approach depends on a set of classifiers that are competent on different regions of the feature space; in other words, they depend on a good performance of the oracle.

Using this abstract fusion model, Table IX shows the upper limit of the classifiers and representations adopted in this work. As we can see, despite of the intrinsic complexity of the problem, the performance of the oracle is very high. Considering a single architecture of classifier trained with six different representations, the upper limit of the system achieves 93.9% in average, except for the QDA classifier that reaches 100% for the subset of 40× magnification images. Considering all the architectures and representations (24 experts), the upper limit increases up to 99% in average. Note that for the both magnification factors of 40× and 400×, all test images could be correctly classified by at least one of the classifiers in the pool.

TABLE IX
SUMMARY OF ACCURACY OF THE ORACLE (%)

Classifier	Magnification factor			
	40×	100×	200×	400×
1-NN	91.5	91.5	93.1	91.5
QDA	100	96.9	96.2	97.7
RF	92.3	91.5	90.8	92.3
SVM	95.4	95.4	94.6	97.7
All classifiers	100	98.5	97.7	100

The first four lines show the oracle for each classifier using six different representations. The last line reports the oracle considering all the 24 classifiers reported in Table VI.

TABLE X
HYPOTHETICAL CONFUSION MATRICES FOR THE ORACLE

	40×		100×		200×		400×	
	B	M	B	M	B	M	B	M
B	1.00	0.00	1.00	0.00	0.88	0.12	1.00	0.00
M	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

B: benign, M: malignant.

Table X presents the hypothetical confusion matrices for the oracle. As we can see, the proposed pool of classifiers is able to solve most of the confusions. The challenge now lies in defining a winner strategy to select the classifiers given an input image.

V. Conclusion

In this paper, we have presented a dataset of BC histopathology images called BreaKHis, which we make available to the scientific community, and a companion protocol (i.e., the folds) for two-class classification of benign versus malignant images. We have performed some first experiments involving six state-of-the-art feature vectors and four classifiers. They have shown room for improvement is left, but also that the complementarity of the magnification factors should be investigated in the future, to design a possible coarse-to-fine strategy for processing the different magnification factor images. One may also consider that different features should be used to describe the different magnification factors. The oracle results also show that a single-classifier might not be enough, and that designing a strategy to combine or select the classifiers given an input image should help to increase the accuracy.

Additional challenges include multiclass classification for both the malignant and the benign image sets. Also, the high false positive rate that we have highlighted in this work may be decreased by implementing a rejection scheme.

By making this dataset available for research purposes, we hope to foster research in computer-aided diagnosis for BC histopathology, and also in ensemble classification by providing a real life, challenging dataset. Future studies may provide some feedback to the pathologist, so as to help him analyzing these images and defining a strategy to identify areas to be explored.

Acknowledgment

The authors would first like to thank the valuable collaboration of P&D Laboratory-Brazil, for providing the slide collection of histopathological studies of breast tissue used in the preparation of the presented image database. In particular, they would like to acknowledge and thank the pathologists F. Negretti and F. S. Rechia of P&D, for their valuable feedback throughout the revision process. Also, they would like to thank C. E. Pokes, a medical student from the State University of West Parana (UNIOESTE), for his great commitment to the digitization process. Finally, the authors would like to thank the reviewers and editors for their very helpful suggestions and insightful comments which have resulted in an improved paper.

References

[1] P. Boyle and B. Levin, Eds., World Cancer Report 2008. Lyon: IARC, 2008. [Online]. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf

[2] J. E. Joy *et al.*, Eds., *Saving Women's Lives: Strategies for Improving Breast Cancer Detection and Diagnosis*. Washington, DC, USA: Natl. Acad. Press, 2005.

[3] B. Stenkvist *et al.*, "Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," *Cancer Res.*, vol. 38, no. 12, pp. 4688–4697, 1978.

[4] Breastcancer.org. (2012) Biopsy. [Online]. Available: http://www.breastcancer.org/symptoms/testing/types/biopsy

[5] R. Rubin *et al.*, Eds., *Rubin's Pathology Clinicopathologic Foundations of Medicine*, 6th ed., Philadelphia, PA, USA: Williams & Wilkins, 2012.

[6] S. R. Lakhani *et al.*, *WHO Classification of Tumours of the Breast*, 4th ed. Lyon, France: WHO Press, 2012.

[7] M. N. Gurcan *et al.*, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.

[8] C. Desir *et al.*, "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2677–2683, Sep. 2012.

[9] M. Kowal *et al.*, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1563–1572, 2013.

[10] P. Filipczuk *et al.*, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2169–2178, Dec. 2013.

[11] Y. M. George *et al.*, "Remote computer-aided breast cancer detection and diagnosis system based on cytological images," *IEEE Syst. J.*, vol. 8, no. 3, pp. 949–964, Sep. 2014.

[12] Y. Zhang *et al.*, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," *Mach. Vision Appl.*, vol. 24, no. 7, pp. 1405–1420, 2013.

[13] Y. Zhang *et al.*, "One-class kernel subspace ensemble for medical image classification," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 17, pp. 1–13, 2014.

[14] M. Veta *et al.*, "Breast cancer histopathology image analysis: A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, May 2014.

[15] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[16] A. L. Mescher, *Junqueiras Basic Histology: Text and Atlas*. New York, NY, USA: McGraw-Hill, 2013.

[17] T. Ojala *et al.*, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[18] Z. Guo *et al.*, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.

[19] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. 3rd Int. Conf. Image Signal Process.*, 2008, vol. 5099, pp. 236–243.

[20] R. Haralick *et al.*, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[21] N. A. Hamilton, *et al.* (2007). Fast automated cell phenotype image classification. *BMC Bioinformatics*. *8*. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/110

[22] E. Rublee *et al.*, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2564–2571.

[23] J. Martins *et al.*, "Forest species recognition based on dynamic classifier selection and dissimilarity feature vector representation," *Mach. Vision Appl.*, vol. 26, no. 2, pp. 279–293, 2015.

[24] J. Paivarinta *et al.*, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Proc. 17th Scandinavian Conf. Image Anal.*, 2011, pp. 360–369.

[25] L. P. Coelho *et al.*, "Structured literature image finder: extracting information from text and images in biomedical literature," in *Linking Literature, Information, and Knowledge for Biology* (ser. LNCS) vol. 6004, C. Blaschke and H. Shatkay, Eds. New York, NY, USA: Springer, 2010, pp. 23–32.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep. 1999, vol. 2, pp. 1150–1157.

[27] H. Bay *et al.*, "Surf: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vision*, May 2006, pp. 404–417.

[28] E. Rosten *et al.*, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, Jan. 2010.

[29] M. Calonder *et al.*, "BRIEF:binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 778–792.

[30] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000, vol. 25(11), pp. 120–125.

[31] C. Cortes and V. Vapnik, "Suport-vector networks," *Mach. Learning*, vol. 20, pp. 273–297, 1995.

[32] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learning Res.*, vol. 12, pp. 2825–2830, 2011.

[34] T. Fawcett, "An introduction to ROC analysis," *Pattern Recog. Lett.*, vol. 27, pp. 861–874, 2006.

[35] J. Kittler *et al.*, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[36] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. New York, NY, USA: Wiley, 2014.

[37] A. S. Britto Jr., R. Sabourin, and L. S. Oliveira, "Dynamic selection of classifiers—A comprehensive review," *Pattern Recognitional*, vol. 47, no. 11, pp. 3665–3680, 2014.

Authors' photographs and biographies not available at the time of publication.