BreakHis based Breast Cancer Automatic Diagnosis using Deep Learning: Taxonomy, Survey and Insights

Article *in* Neurocomputing · September 2019

DOI: 10.1016/j.neucom.2019.09.044

4 authors:

Yassir Benhammou
University of Granada
9 PUBLICATIONS **293** CITATIONS
SEE PROFILE

Achchab Boujemâa
Université Hassan 1er and Université Mohammed VI Polytechnique
127 PUBLICATIONS **799** CITATIONS
SEE PROFILE

Siham Tabik
University of Granada
101 PUBLICATIONS **9,465** CITATIONS
SEE PROFILE

Francisco Herrera
University of Granada
1,088 PUBLICATIONS **119,472** CITATIONS
SEE PROFILE

BreakHis based Breast Cancer Automatic Diagnosis using Deep Learning: Taxonomy, Survey and Insights

Yassir Benhammou[a,b], Boujemâa Achchab[b], Francisco Herrera[a], Siham Tabik[a]

[a] *Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain*
[b] *Systems Analysis and Modeling for Decision Support Laboratory, National School of Applied Sciences of Berrechid, Hassan 1st University, Berrechid 218, Morocco*

Abstract

There are several breast cancer datasets for building Computer Aided Diagnosis systems (CADs) using either deep learning or traditional models. However, most of these datasets impose various trade-offs on practitioners related to their availability or inner clinical value. Recently, a public dataset called BreakHis has been released to overcome these limitations. BreakHis is organized into four magnification levels, each image is labeled according to its main category (Benign/Malignant) and its subcategory (A/F/PT/TA/PC/DC/LC/MC). This organization allows practitioners to address this problem either as a binary or a multi-category classification task with either a magnification dependent or independent training approach. In this work, we define a taxonomy that categorize this problem into four different reformulations: Magnification-Specific Binary (MSB), Magnification-Independent Binary (MIB), Magnification-Specific Multi-category (MSM) and Magnification-Independent Multi-category (MIM) classifications. We provide a comprehensive survey of all related works. We identify the best reformulation from clinical and practical standpoints. Finally, we explore for the first time the MIM approach using deep learning and draw the learnt lessons.

Keywords: Breast cancer, BreakHis dataset, Histopathological images, Computer aided diagnosis, Deep learning, Data preprocessing

Email addresses: benhammou@correo.ugr.es (Yassir Benhammou), achchab@estb.ac.ma (Boujemâa Achchab), herrera@decsai.ugr.es (Francisco Herrera), siham@ugr.es (Siham Tabik)

1. Introduction

In spite of the massive growth in breast cancer incidence during last years, its death rate has considerably decreased [1]. This drop in mortality incidence has mainly occurred in developed countries. Especially, those achieving important breakthroughs in early detection methods through medical images analysis [2]. Generally, the most infallible breast cancer diagnosis is histopathological biopsy examination [3]. The latter is carried out by pathologists using fine needle expelled slides from breast tissue. For each patient, a number of breast tissue slides are analyzed with various microscopic magnification levels to better highlight Regions of Interest (ROIs). Nevertheless, pathologist's interpretation is often deviated by several human factors such as eye fatigue, in addition to device-dependant influences.

To alleviate the risk of putting a patient's life at stake, domain experts entrust their assistance in this task to Computer Aided Diagnosis (CAD) systems [4, 5, 6, 7]; in which a large research community is working on its improvement. The golden source for these systems is data collected and annotated by experts during similar real decision-making situations. There are mainly three histopathological breast cancer diagnosis datasets; Bioimaging[1][8], MITOSATYPIA[2][9] and [10], but they may be partially or completely not available. They also suffer from a lack of sufficient clinical value as it can be seen in Table 1:

Dataset	Availability	Main classes	Sub-classes	Target problem	Magnification factors	Image distribution
Bioimaging [8]	only the training set	Carcinoma /non Carcinoma	Carcinoma(In situ carcinoma/ Invasive carcinoma), Non-carcinoma(Normal tissue/ Benign tissue)	Breast cancer Classification	x200	120 images 20 images (not available)
MITOSATYPIA [9]	only the training set	nuclear atypia score(1/2/3)	—	Detection of mitosis, counting mitosis and evaluation of nuclear atypia score	x20 x40	284 images at x20 and 1136 images at x40
[10]	No	Nuclei / Non-nuclei	—	Nuclei classification	x20	37 images

Table 1: Characteristics of the three existent breast cancer histopathological datasets.

To our knowledge, the most recent public breast cancer histopathological dataset which has a higher clinical value than previously mentioned ones is BreakHis[3][11]. The latter is organized into four magnification levels, $\times 40$,

[1] http://www.bioimaging2015.ineb.up.pt/dataset.html
[2] https://mitos-atypia-14.grand-challenge.org
[3] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

×100, ×200, ×400. Each image is labeled as either benign or malignant categories, and also as one of these eight sub-categories: Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenoma (TA) for benign images and Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC) and Papillary Carcinoma (PC) for malignant ones. These characteristics allows practitioners to reformulate this problem as different classification tasks: a binary classification to predict whether an input image is benign/malignant, or a multi-category classification to predict its specific sub-category among (A/F/TA/PT/DC/LC/MC/PC) sub-classes. In each classification task, the adopted model use either a magnification-specific or a magnification-independent training scenario. In the magnification-specific approach, one model is trained on each magnification level subset, resulting in four specific models. While in the magnification-independent approach, a unique model is trained using all magnification levels combined.

Since its release, over 45 studies evaluated the potential of BreakHis. In this paper we analyzed BreakHis from four different perspectives:

- We proposed a taxonomy that categorizes BreakHis related works into four different groups: MSB, MIB, MSM and MIM.
- We provided a comprehensive survey of all pre-processing, processing and post-processing methods proposed in different related works. These works were organized using the proposed taxonomy.
- We identified the best reformulation from a clinical as well as a practical standpoint by analyzing different aspects used to define our taxonomy. We found that MIM reformulation is the most appropriate one for this problem.
- As the MIM approach has not been explored yet in literature, we evaluated its potential for the first time using deep learning.

This paper is organized as follows: section 2 presents BreakHis dataset and its characteristics. In section 3, we present our proposed taxonomy. Section 4, 5 and 6 present a complete overview of all BreakHis related works that have adopted MSB, MIB and MSM respectively. In section 7, we identify the best reformulation from a clinical as well as a practical standpoint. In section 8, we evaluate the MIM reformulation with BreakHis dataset for the first time and report the learnt lessons. Finally, we conclude the paper in section 9.

2. BreakHis histopathological breast cancer dataset

This section provides a complete description of BreakHis dataset.In particular:

- Section 2.1 describes the main characteristics, distribution and organization of BreakHis
- Section 2.2 analyzes BreakHis limitations in terms of data imbalance and noisy labels
- Section 2.3 presents the experimental protocol established by BreakHis authors for CAD system development and evaluation

2.1. BreakHis dataset description

Alike most cellular pathologies, breast tumors examination relies essentially on histopathological slides [4]. The latter are surgically expelled tissues from patient's breast after a biopsy operation on the ROI. BreakHis current version is composed of 7909 histopathological biopsy images taken from 82 patients. These images were collected by P&D Laboratory in Brazil from January 2014 to December 2014. BreakHis is divided into two main malignancy classes: benign and malignant, with 2480 benign and 5429 malignant tumor images. Each malignancy class is distributed into four different subcategories based on the tumor appearance under microscope. Each patient in this dataset has a number of images annotated with the main and subcategory classes. A summary of image and patient distributions over main classes and different sub-categories is presented in Table 2.

Main category	**Benign**				**Total**	**Malignant**				**Total**	**Total**
Sub-category	**A**	**F**	**PT**	**TA**	**Benign**	**DC**	**LC**	**MC**	**PC**	**Malignant**	**both**
Number of images	444	1014	453	569	2480	3451	626	792	560	5429	7909
Number of patients	4	10	3	7	24	38	5	9	6	58	82

Table 2: Image and patient distribution among the main categories and each sub-category.

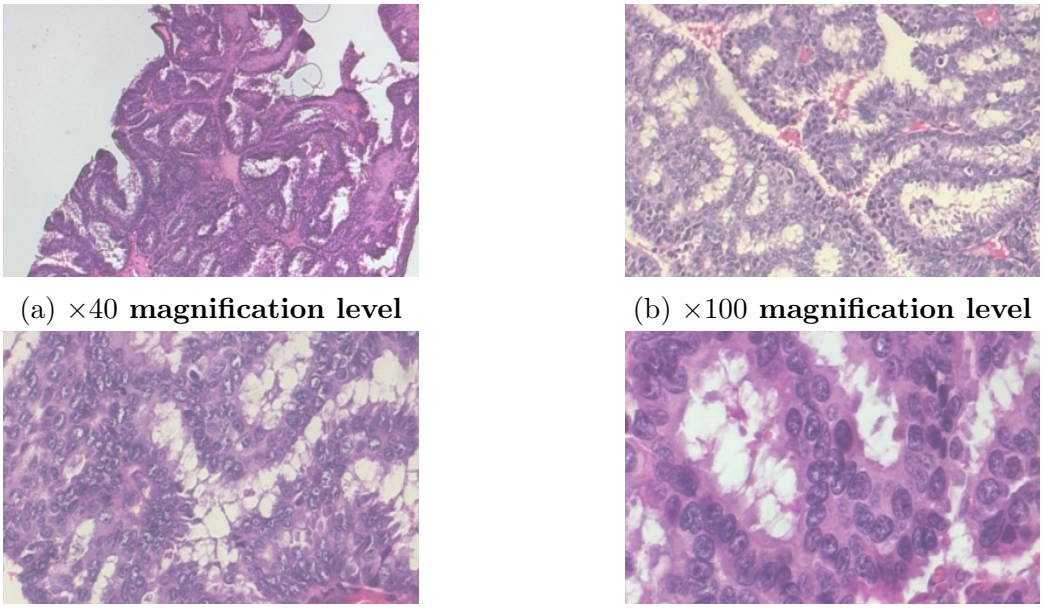
BreakHis images were collected using the same clinical process adopted in similar histopathological datasets [12].In general, for each patient, several breast tissue samples are aspired with a fine biopsy needle in the operating room. Then, each sample undergoes the following preparation phases: First, formalin fixation and embedding in paraffin to preserve the original tissue structure and its molecular composition. Then, $3\mu m$ thickness sections are

extracted from paraffin outcomes using a high precision microtome. Afterwards, these sections are mounted on covered glass slides for visualization under microscope.

Often, components of interest such as nuclei or cytoplasm are not clearly visible in raw tissue sections. Hence, an essential operation called tissue staining takes place before visualization. This staining step aims to highlight each component for a better morphological analysis under the microscope. Several staining methods exists, and the most used one is Hematoxylin and Eosin (H&E) [4]. In BreakHis, Hematoxylin(H) binds to DNA and thereby dyes the related structures of interest (i.e. in most cases nuclei) with blue/purple, and Eosin(E) binds to proteins and dyes other structures including cytoplasm and stroma with pink.

Few years ago, this clinical workflow was concluded by sending to pathologists all stained-slides in physical version for microscopic analysis. Nowadays, with the appearance of Whole Slide Image (WSI) scanners, a slide digitization step is added on the top of this process, and a digitized version of these slides is also sent to pathologists, then annotated and stored in the laboratory information system. This additional step may also include slide manipulation to collect slides with different magnification factors. Notably, BreakHis images were stored with four magnification levels ($\times$40, $\times$100, $\times$200, $\times$400). During analysis and annotation, pathologists starts by identifying ROIs in the lowest magnification level slide ($\times$40), then dives deeper in the latter using higher magnification levels ($\times$100, $\times$200) until having a profound insight ($\times$400). To illustrate this process, a BreakHis slide sample captured with four different magnification factors is presented in figure 1.

BreakHis distribution into four magnification levels for each tumor category and sub-category is presented in Table 3. Each magnification subset contains a different number of images. This number is ranging from 1794 images in $\times$400 subset to 2051 images in $\times$100 subset. For patient distribution, each magnification factor subset contains exactly 82 patients because each patient has images taken with all magnifications. After further statistical exploration, we found that each magnification factor subset contains around 24 images per patient. In average 24, 25, 24, and 22 images per patient are available in $\times$40, $\times$100, $\times$200, and $\times$400 subsets respectively.



(a) ×40 **magnification level** (b) ×100 **magnification level**

(c) ×200 **magnification level** (d) ×400 **magnification level**

Figure 1: The same malignant PC tissue captured with different optical magnification levels(×40, ×100, ×200, ×400), taken from the patient with ID:9146 in BreakHis dataset

Main category		**Benign**				**Total**	**Malignant**				**Total**	**Total**
Sub-category		**A**	**F**	**TA**	**PT**	**Benign**	**DC**	**LC**	**MC**	**PC**	**Malignant**	**of both**
Number of images at each magnification level	×40	114	253	109	149	598	864	156	205	145	1370	1968
	×100	113	260	121	150	614	903	170	222	142	1437	2051
	×200	111	264	108	140	594	896	163	196	135	1390	1984
	×400	106	237	115	130	562	788	137	169	138	1232	1794

Table 3: The distribution of BreakHis images into four magnification levels for both main tumor categories and each sub-category.

2.2. BreakHis data imbalance and noise labels

BreakHis histopathological dataset has been mainly released to overcome the lack of previous datasets in terms of availability and clinical content richness. Nevertheless, BreakHis still has some common abnormalities that are present in all medical datasets due to this disease nature and limitations in medical data acquisition techniques. Particularly, the following characteristics are the most relevant in BreakHis, and should be taken into consideration when building a robust CAD system:

Data imbalance As it can be seen in Table 2 and 3, BreakHis data imbalance occurs at different levels. The Imbalance Ratio (IR) between malignant and benign classes is 0.41 at patient level and 0.45 at image level. An uneven distribution is also present between different sub-categories at image and patient levels. This data imbalance issue could bias the discriminative capability of a CAD system towards the majority class at image and patient levels during binary and multi-category classification tasks.
Label noise During its evolution, breast tumor expands gradually from a region to another in breast tissue. Hence, sometimes images captured from the same breast tissue may contain different regions with various breast cancer stages, resulting in different sub-category annotations for the same image. Images corresponding to the patient with ID:13412 are replicated in two malignant sub-categories: (DC) and (LC). This special case can confuse the CAD model during a multi-category classification task that tries to learn discriminative features between both malignant sub-classes.

2.3. BreakHis experimental protocol

BreakHis has been proposed with the aim to constitute a benchmark for breast cancer CAD systems. Therefore, its authors proposed the following unified experimentation protocol in the MSB reformulation:
For each magnification factor subset, five evaluations are performed.
During each evaluation the used magnification subset is randomly divided into 70% for training and 30% for test.
To guarantee that the CAD model generalizes to unseen patients, the patients used to build the training set are not used for test.

For a fair comparison between different CAD systems, BreakHis authors proposed two classification level metrics:

- **Image Level Accuracy** This first metric does not take into account any patient information. It is the standard classification accuracy at image level defined as:

$$ILA = \frac{I_{corr}}{I_{tot}} \tag{1}$$

  Where I_{tot} is the total number of test images and I_{corr} is the number of correctly classified test images.

- **Patient Level Accuracy** The second metric reflects the achieved performance in a patient-wise manner. Firstly, an individual score is computed for each single patient P_i, and then the Patient Level Accuracy

(PLA) is calculated as an average over all test patients scores.
The patient score for each patient P_i is:

$$Score(P_i) = \frac{I_{Pi_{corr}}}{I_{Pi_{tot}}} \tag{2}$$

Where $I_{Pi_{corr}}$ and $I_{Pi_{tot}}$ are respectively the number of correctly classified test images and the total number of test images for a particular patient P_i.
Afterwards, PLA is measured by:

$$PLA = \frac{\Sigma_{i=1}^{i=N}(Score(P_i))}{N} \tag{3}$$

Where N is the number of patients in the test set.

- **Additional metrics** Conventionally, during cancer diagnosis a malignant case is considered as positive while a benign one is considered as negative. The sensitivity of a CAD system towards positive cases (malignant) is clinically more important. The first two metrics ILA and PLA are based on the standard accuracy which is calculated using the number of correct predictions in both negative and positive classes similarly. Therefore, some authors used other evaluation metrics such as $F1_{score}$ or Area Under Curve (AUC) because they are directly related to the CAD's sensitivity towards positive cases.

  - $F1_{score}$ [13], also called $F1_{measure}$ or F_{score}, is adopted to better highlight a CAD system's sensitivity to (positive) malignant cases as they are of high interest in this kind of medical diagnosis. F1-score is defined as the harmonic mean between sensitivity (also called Recall) and Precision:

$$F1_{score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

    Where:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5}$$

$$Recall = Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{6}$$

– Area Under Curve (AUC) [13] is adopted in other works to measure how the sensitivity (recall) and false positive rate trade off, so in that sense it is already a broader metric. More importantly, unlike standard accuracy, AUC is not a function of threshold. It allows an evaluation of the classifier over all possible threshold values.

3. Reformulations Taxonomy

Over the past three years, several BreakHis based CAD systems have been built. To survey all this research we proposed a new taxonomy that organizes BreakHis related works into four different groups according to their adopted reformulation of this classification problem. The main motivation behind the proposed taxonomy, is to understand all the proposals that addressed BreakHis dataset for building CAD systems, their strengths and their weaknesses. In addition, this categorization conducted us to identify and evaluate the most suitable reformulation for this problem from clinical as well as practical standpoints, which to our knowledge has never been explored or even identified before in the literature. For instance, we present in this section the adopted taxonomy and its four groups (see Table 4).

Related works:	**Reformulated as**	**Number of works**
[11][14][15][16][17][18][19] [20][21][22][23][24][25][26] [27][28][29][30][31][32][33] [34][35][36][37][38][39][40] [41][42][43][44][45][46][47] [48][49][50][51][52]	MSB	40
[53][54]	MIB	2
[55][56][57][58]	MSM	4
Explored in this paper for the first time	MIM	-

Table 4: BreakHis related works and their corresponding reformulations in our taxonomy.

Where MSB reformulation:classifies the input image as benign or malignant depending on its magnification factor. MIB reformulation:classifies the

input image as benign or malignant regardless of its magnification factor. MSM reformulation:classifies the input image into one of the eight subcategories with taking into consideration its magnification factor. MIM reformulation:classifies each image into one of the eight subcategories regardless of its magnification factor. To illustrate the proposed taxonomy, we present in figure 2 the inputs, classifiers and outcomes of each reformulation:

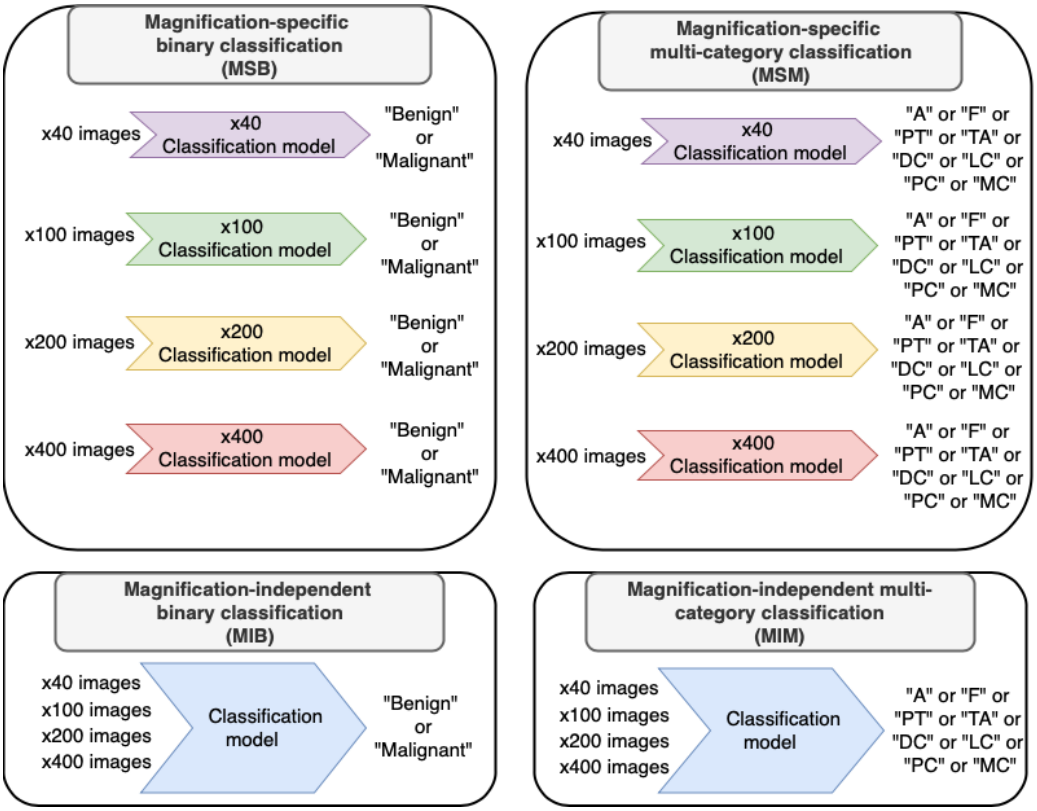


Figure 2: An illustration of each reformulation in the proposed taxonomy

MSB, MIB and MSM works will be presented in sections 4, 5 and 6 respectively, while MIM will be explored for the first time in this paper in section 8. A work that belongs to two different groups will be reported only within the group that represents its main reformulation, while the results of its both reformulations will be included separately in their corresponding tables. Namely, MSM is the main reformulation of [55, 56, 57] while MSB is their secondary reformulation. Besides, the main reformulation of [15] is MSB while its secondary reformulation is MSM.

4. Magnification-specific binary classification works

In this section, we will report all works that used an MSB approach. In fact, 85% of BreakHis CAD systems adopted this reformulation. These MSB works will be organized as follows:

- Section 4.1 summarizes works that adopted a traditional handcrafted based model
- Section 4.2 presents works that used deep learning based models
- Section 4.3 covers works that brought their contributions to the pre-processing phase
- Section 4.4 describes works that developed their CAD systems with a content-based histopathological image retrieval approach
- Section 4.5 highlights works that focused on domain adaptation

Then, in Table 5 we will summarize all these works and for each one we will present its best results, the adopted pre- and post-processing approaches, the used model and the learning strategies. Depending on the experimental setup used in each paper, results are going to be presented either as a mean value with a standard deviation over various trials or as a unique trial value. Depending on their availability, results are reported at different metrics levels including PLA, ILA, AUC and F1-score. For the seek of space, we used an abbreviation for each method, and the corresponding dictionary can be found in Appendix A.

4.1. Handcrafted descriptors based models

First BreakHis CAD systems adopted a traditional dual-stage approach, by extracting handcrafted features from the images, then using them to train a standalone classifier. At features extraction phase, some works evaluated multiple descriptors with the aim to select the most representative ones, while others used an unique descriptor. In this part, we will present all these works with their adopted image descriptors:

First, BreakHis authors explored in [11] the effectiveness of six state-of-the-art handcrafted features descriptors: Local Binary Patterns (LBP)[59], its variant Completed Local Binary Pattern (CLBP) [60], Local Phase Quantization (LPQ) [61], Gray Level Co-Occurrence Matrices (GLCM) [62], Parameter-Free Threshold Adjacency Statistics (PFTAS) [63] and Oriented FAST and

Rotated BRIEF (ORB) [64], associated with four different classifiers: 1-Nearest Neighbor (1-NN) [65], Quadratic Linear Analysis (QDA)[66], Support Vector Machines (SVM)[67], and Random Forests of decision trees [68]. Then, To evaluate the effectiveness of fractal dimension [69] as the only descriptor; authors in [15] trained an SVM classifier with the fractal dimension of each image. Results of which demonstrated that using fractal dimension as a unique descriptor is more suitable when classifying $\times 40$ images with a lot of self-similarities, but meaningless in higher magnification images with less self-similarities. As an additional part of this work, a multi-category classification task was elaborated with 16 experiments, each one classifying a benign and a malignant sub-classes.
Afterwards, authors in [24] evaluated different handcrafted descriptors in conjunction with a k-NN classifier, including: LBP, GLCM, PWT) and TWT. After that, authors in [33] proposed to use Zernaike moment, image entropy and fractal dimension features through a signal processing multilevel iterative variational mode decomposition (VMD)[70], and select the most relevant features using Relief [71], before classifying them with a Least squares support vector machine (LS SVM). In [18], authors tried to enable an L1-norm sparse SVM (SSVM) [72] to select the most relevant features from BreakHis images. According to their work, L1-norm is inconsistent in establishing features selection precisely and the SSVM could be biased towards large hyper-plane coefficients. To enhance its features selection quality, they assigned a weight to each feature depending on its Wilcoxon rank sum[73]. Lately, authors in [39] evaluated KAZE features [74] performance in a bag-of-features approach. Then, transformed these features into histogram information using an approximate Nearest Neighbour algorithm, and used them to train an SVM classifier.
Limitations Results achieved by different traditional handcrafted features were considered relatively acceptable as preliminary results but highly unstable. In fact, the major drawback of these traditional approaches is that: the model's quality depends on the extracted features, while acquiring a highly representative features is a very complex task. One of the main difficulties is the right descriptor choice, and even when various descriptors are combined together to increase their discriminative power, or post-transformed to select the most appropriate ones, their achieved results remains relatively low and unstable between different magnification levels.

4.2. Deep learning based models

To mitigate traditional practices limitations, some authors thought of entrusting features extraction as well as classification tasks to deep learning models giving their ability to select directly the most significant global features. These researches tried to successively boost their deep learning based approach, starting with the exploration of various models, to the analysis of different learning and adaptation strategies. In this part, we will organize these works according to their contributions at different deep learning aspects:

From traditional towards deep learning approach BreakHis authors were the first researchers to move towards the evaluation of a deep learning based CAD system for this dataset, by entrusting features extraction and classification tasks to a CNN model in [14]. They started by evaluating LeNet [75], but its results were lower than those reported by their previous traditional based model in [11]. Therefore, they opted for AlexNet [76] as a relatively deeper network.

Deep learning models trained with handcrafted features Generally, CNN models are provided with raw images as input. However, authors [35] were convinced by the importance of textural and pixel distributions contained in handcrafted features such as LBP or histogram descriptors. By consequence, they evaluated CNNs provided with various handcrafted features in comparison to those provided with raw images. After exploring different combinations, their best results were achieved with a model called (Model1-CNN-CH). The latter was a CNN model with residual blocks inspired from ResNet [77], and provided with a concatenation of extracted local-features using Contourlet Transform (CT)[78] and Histogram information descriptors.

CNN models comparison To find the adequate CNN for this classification task, authors in [30] compared the performance of three different CNN models: CaffeNet which is an AlexNet variant, GoogleNet [79] and ResNet-50. Results of which proved the efficiency of ResNet, the necessity of data augmentation, fine-tuning all layers, providing this CNN with large WSIs instead of small patches, and using ensemble learning by combining different magnification-specific models. In our last work [40], we explored the performance of another CNN which is a GoogleNet variant called Inception-v3 [80], and our results shown the efficiency of this CNN in comparison to shallower ones used in previous works.

Pre-trained CNN for features extraction use The improvement brought by the first CNN evaluations in BreakHis [14], encouraged its authors to explore further deep learning capabilities. They evaluated in [16] a transfer learning strategy with a pre-trained AlexNet and DeCAF features extraction approach [81]. The latter consists of extracting features from the pre-trained AlexNet's last layers, then using them to train a standalone classifier. Afterwards, authors in [31] explored the impact of three different dimensionality reduction methods on features extracted from a pre-trained VGG [82]: Principal Component Analysis(PCA)[83], Gaussian Random Projection(GPR)[84] and Correlation-Based Feature Selection(CBFS)[85].
Generally, when a pre-trained CNN is adopted as a features extractor, only its final layers features are exploited. To evaluate the potential contained in every layer of a pre-trained DenseNet-169 [86], in conjunction with XGBoost classifier [87], authors in [42] proposed a sequential features extraction framework. This evaluation proved that the last convolutional layers provide more significant features than the final fully connected layers. Another interesting finding of this work is that: lower level layers contribute significantly to $\times 40$ images classification. Similarly, mid range magnifications $\times 100$ and $\times 200$ are better represented by mid level features. While $\times 400$ images are better captured by higher level layers.
Pre-trained CNN for Fine-tuning use For fine-tuning, various practices are adopted. In some works, all the pre-trained CNN layers are fine-tuned; while in others, only the last fully connected layers are retrained. Authors in [29] proposed a dual-stage fine-tuning method, which retrained only the fully connected layers in a first place, then the whole network. To justify this choice they also evaluated it against each one of its two independent stages.
Features extraction versus fine-tuning use In [45], the authors demonstrated that fine-tuning the three last layers of a pre-trained AlexNet is more efficient than SVM classification of concatenated features extracted from two pre-trained CNNs (AlexNet and VGG16).
Features extraction combined with fine-tuning An ImageNet pre-trained CNN extracted features are meant to be the common high level features between ImageNet and BreakHis classification tasks. However, the used CNN is not supervised to extract necessary features for BreakHis classification. Thus, a gap is generated between the extracted features and the required specific domain features [88]. To mitigate this gap, [37] proposed a hybrid transfer learning approach called deep domain knowledge-based features model, by adding a preliminary knowledge adaptation step that consists of retraining

(fine-tuning) the pre-trained CNN on BreakHis classification task in a first place for more efficient features extraction.

CNN Features extraction versus handcrafted features extraction Authors in [47] evaluated features obtained with traditional handcrafted descriptors in comparison with those extracted from a pre-trained AlexNet. Surprisingly, LBP handcrafted features have proven to be slightly better than AlexNet features. However, this comparison remains very restrictive with a relatively shallow CNN which was found also in [14] to be barely capable of outperforming handcrafted based models.

CNN features post-encoding using Fisher Vector Post-encoded CNN features using Fisher Vector (FV)[89] are known for their good classification potential [90, 91]. To evaluate their performance in BreakHis problem, authors in [17] started with fine-tuning an pre-trained VGG model as a preliminary adaptation of this CNN on BreakHis classification. Then extracted a dense set of local features from its last convolutional layer in order to encode them into FV descriptor. Afterwards, the same authors presented a second work [21] with the aim to overcome FV high dimensionality issue. In fact, they proposed a supervised intra-embedding model designed to embed each block of the FV descriptor into a lower dimensional feature space, using a dimensionality reduction algorithm based on a multilayer neural network. In their next work [92], they proposed new features representation method called Component Selective Encoding (CSE). Therefore, they used the same pre-trained VGG as in [17], along with a new adapted dimensionality reduction method inspired from one of their previous works [93]. The latter aims to reduce each FV component individually, unlike [21] where they were reduced uniformly. This adaptation was justified by the fact that some regions in these images are more relevant than others.

CNN architecture adaptation Inspired by the effectiveness of GoogleNet inception module in capturing multi-scale features with different convolutions, Authors in [23], proposed a new version of this module called "Transition module" and integrated it to AlexNet. This new version was designed with the aim to ease the abrupt transition between the last convolution layer and the first fully connected layer. Unlike inception module, no prior dimensionality reduction was included to this transition module. Then, authors in [26] proposed a new CNN architecture composed of fifty convolutions, and compared it to several handcrafted features based models. Afterwards, authors in [20] designed a new CNN called BiCNN inspired from GoogleNet architecture. To use this CNN in the binary classification task they proposed

to take into consideration sub-class information in conjunction with binary labels of each image as a prior knowledge. They claimed that this consideration of both annotation levels could help the proposed model to better learn features distance between binary classes. In [34], the authors tried to leverage recent findings in rotation equivariant CNNs [94] with the inherent symmetry under rotation and reflection of histopathological images, in order to build a rotation and reflection equivariant CNN inspired from DenseNet. Lately, another work [50] proposed a CNN model composed of different layers combinations (convolutional, pooling and fully connected layers), whereas various compositions and hyper-parameters were evaluated to determine the most adequate architecture for BreakHis classification.

Multiple Instance Learning with CNN Practically, CNNs requires original WSIs images to be resized to fit in their input layer. Some practitioners prefer to extract low size patches from original WSIs in order to avoid loosing any discriminative information. However, adopting a patch based approach is very challenging, since only a small number of extracted patches is correctly labeled. This fact is due to the presence of benign areas in malignant WSIs. To address this mislabeled patches issue, authors in [36] proposed to use a Multiple Instance Learning (MIL) approach with randomly extracted 64×64 patches. In fact, they noticed that BreakHis distribution at patient and image level is similar to MIL reasoning, and adapted its formulation to two different settings. The first one meets the labeling at image level, where each image was considered as a bag of instances. The second setting considered each patient as a bag. They explored twelve different MIL methods including recent ones such as deep learning based MIL-CNN [95] and non-parametric MIL [96]. In another MIL based approach [49], the authors introduced a new MIL CNN layer termed multiple instance pooling (MIP) layer with the aim to select from each bag their most discriminative instances with the higher feature responses, instead of capturing all their instances. This constraint was integrated into the loss function by considering only the loss associated to higher activation instances.

Deep active learning To avoid mislabeled patches when adopting a patch based approach, practitioners are forced to annotate all extracted patches. However, this task is expensive, very tedious and time-consuming. In order to reduce this labeling burden, authors in [44] proposed a deep active learning framework enhanced with a boosted confidence approach. This approach is based on an active learning model which is firstly initialised with very limited labeled data. Then, it selects at each iteration the lowest confidence

unlabeled samples (highest entropy samples) and give them to domain experts for annotation. By consequence, it reduces considerably the annotation cost. However, it ignores the less representative samples (higher confidence samples with lower entropy) and their potential. Therefore, authors in [44] provided these remaining samples to the model itself for auto-annotation without any additional manual-annotation cost.

Ensemble learning To prove the effectiveness of ensemble learning with features learned by different classifiers at various scales, authors in [19] explored the performance of an ensemble of different magnification-specific model where each one is a pre-trained GoogleNet. In fact, these CNNs were trained separately in a magnification-specific way, but for each test image an ensemble of all these magnification-specific CNNs was aggregated using a majority voting rule.

Deep Belief Network Inspired by the outstanding results achieved with Deep Belief Networks (DBN) in many fields applications [97], researchers in [48] used a Deep Belief Network (DBN) composed of four stacked Restricted Boltzmann Machine (RBM). But, instead of using raw images they provided it with handcrafted Tamura features [98].

Autoencoder In [43], the authors proposed a hybrid framework that starts with a LandMark ISOMAP (L-ISOMAP) embedding [99] to extract the most significant features in BreakHis images, followed by an SSAE with two stacked sparse Autoencoders and a classification output layer. This choice was motivated by the fact that Autoencoders have shown distinguishable results in different image classification tasks [100]. In their results, they stated that this approach allowed them to achieve an improvement in classification rate while reducing the overall computational cost.

Limitations Results achieved by different deep learning models are considerably higher than those presented with traditional approaches. Nevertheless, deep learning models are extremely data-hungry and require a large amount of data, while medical applications such as breast cancer diagnosis always suffer from a lack of data. To mitigate this limitation, often practitioners are forced to adopt artificial data augmenters as a preprocessing. In addition, determining the best hyper-parameters for this kind of models is a black art with no guiding theory. Moreover, unlike handcrafted engineered models where what is learned is easy to comprehend, deep learning approaches are not able to give users feedback or interpretability on the discriminative features used to decide about each patient diagnosis. Besides breast cancer, many deep learning based medical applications

exists and each one of them has different limitations [101]. For a full review on this topic, we refer the reader to the following recent references [102, 103, 104, 105, 106, 107, 108, 109, 110]

4.3. Preprocessing methods

To build these CAD systems either with deep learning or traditional approaches, raw images need various preprocessing transformations. In this part we will present works where the main contribution is related to the preprocessing phase:

Data augmentation Deep learning models, and especially CNNs requires an important volume of training data and particularly when they are to be fine-tuned. Thus, to generate a sufficient number of data samples for fine-tuning a pre-trained inception v3 on $\times 40$ images classification, authors in [28] evaluated different data augmentation techniques and reported their results.

Clustering as a preprocessing To explore the hidden similarities in morphological textures of BreakHis images, authors in [38] adopted a clustering algorithm as a preprocessing step. This method aims to extract statistical and geometrical clusters hidden in the data structure. In order to prove the efficiency of their approach, they evaluated the performance of a CNN provided with these cluster-transformed images using various clustering algorithms in comparison to a CNN provided with raw images. In [27], authors evaluated a segmentation preprocessing step based on a clustering algorithm to highlight nuclei regions in each image before extracting their features and provide them to different classifiers. Lately, in [41], the authors used a K-means clustering on each image to highlight its nuclei segments, before extracting entropy features from these cluster-transformed images using Discrete Wavelet Transform (DWT)[111]. Then, evaluated an SVM classifier trained with these features.

Stain normalization To address the stain variability of BreakHis images, authors in [22] were motivated by learning the color-texture variation of these images instead of reducing the color-variations between them. Therefore, they explored several combinations of various color-texture descriptors along with different classifiers. After identifying the best performing features-classifier combination in each magnifications subset, they combined them in an integrated model. Then, the same authors tried in [25] to provide indications about whether it is possible for a model to learn this color-texture variability instead of normalizing it. Their experiments found that: on one hand; stain normalization could be substituted by joint color-texture features

learning to achieve higher results, on the other hand: gray scale transformation is not a good stain normalization method and could decrease the classification accuracy. Recently, authors in [52] started from the conviction that conventional normalization techniques amplify the existing noise in images when applied directly, and propose a normalization strategy that includes a noise amplification control step.

4.4. Content-based histopathological image retrieval CAD systems

Unlike standard CAD systems, in a Content-based histopathological image retrieval (CBHIR) system, the model search in the dataset for other images with similar content to the query one, and return them to the pathologist to use their diagnostic information as reference. In this part we will present works that adopted a CBHIR system:

One of the first attempts to employ a CBHIR approach is [32]. Authors in this work were aware of the fact that a CBHIR system is computationally intensive especially when retrieving from a large-scale dataset, because it is essentially based on feature vectors comparison to measure images similarities. To address this aspect, they proposed a three-steps framework: Firstly, all images in the dataset undergone a binarization encoding with different tile sizes in order to decrease the required computations and ensure their size-scalabality. Then, for each query image a similarity-based search is conducted to look for the closest proposals in the encoded dataset. Finally, the retrieved images are ranked, giving their similarities to the query image using Hamming distance. To reduce retrieving time, authors in [51] used a pre-trained VGG-19 that extracts class-specific and patient-specific tumorous descriptor simultaneously. In fact, they divided the adopted framework into two different phases. In the first phase, the pre-trained VGG-19 extracts the last layer features as a 1000-dimensional vector from each image in the gallery set (the training set) and then uses these features to train a multi-patient classifier which gives for each image a score evaluating if it belongs to a given patient (82-dimensional vector) and a binary malignancy classifier to determine either is benign or malignant (2-dimensional vector). In the second phase, they extracted from each query image its 1000-dimensional features using a pre-trained VGG-19, then these features are passed through the fine-tuned multi-patient and binary classifiers to find a conjoint patient/class 2-dimensional vector. This vector is used to retrieve similar images from the gallery set.

4.5. Domain adaptation approach

Mostly, all BreakHis models are built with the assumption that the distribution of training and testing data are the same, whereas; others claimed that this assumption is not correct since histopathological images are prepared and stained in different laboratories with different standards, which could adversely degrade the classification rate. In this part we will present works that proposed a domain adaptation approach:

Authors in [46] were the first and only to address the aforementioned issue; they proposed a new learning framework with an unsupervised domain adaptation approach based on the data representation-learning. The goal of this domain adaptation is to reduce the differences between the marginal distributions of source and target domains while learning a new representation for both domains. This method is based on the creation of an invariant space where training and test sets are projected to adapt their different domains.

Work	Preprocessing	Patch/Slide	Features extractor	Classifier	Transfer learning	Training/Test	Metrics	Results(%)			
								×40	×100	×200	×400
[11]	None	WSI	PFTAS	QDA	None	70 % / 30 %	PLA	83.8 ± 4.1	82.1 ± 4.9	84.2 ± 4.1	82.0 ± 5.9
[14]	-Res(350x230) -SMI	-Rnd(32x32, 64x64) -SW(32x32, 64x64)	Emax(AlexNet)		ImageNet	70 % / 30 %	ILA	85.6 ± 4.8	83.5 ± 4.8	84.6 ± 4.2	86.1 ± 6.2
							PLA	90.0 ± 6.7	88.4 ± 4.8	84.6 ± 4.2	86.1 ± 6.2
[15]	Binarization	WSI	FD	SVM	None	50 % / 50 %	F1	97.9	16.5	16.5	25.3
[16]	None	MKV(1,4,6)	CaffeNet	LR	ImageNet	70% / 30%	ILA	84.6 ± 2.9	84.8 ± 4.2	84.2 ± 1.7	81.6 ± 3.7
							PLA	84.0 ± 6.9	83.9 ± 5.9	86.3 ± 3.5	82.1 ± 2.4
[57]	-DAB(IV, Rot, Tr, Flip)	WSI	NDCNN(GoogleNet)		ImageNet	50% / 50%	ILA	95.8 ± 3.1	96.9 ± 1.9	96.7 ± 2.0	94.9 ± 2.8
							PLA	97.1 ± 1.5	95.7 ± 2.8	96.5 ± 2.1	95.7 ± 2.2
[17]	None	WSI	FV(VGG)	SVM	ImageNet	70% / 30%	ILA	87.0 ± 2.6	86.2 ± 3.7	85.2 ± 2.1	82.9 ± 3.7
							PLA	90.0 ± 3.2	88.9 ± 5.0	86.9 ± 5.2	86.3 ± 7.0
[18]	None	WSI	DR(ASSVM, WRS)		None	70% / 30%	PLA	94.97	93.62	94.54	94.42
[19]	-Res(370x230) -DAB(Rot, Flip)	Rnd(224x224)	GoogleNet		ImageNet	80% / 20%	ILA	94.82	94.38	94.67	93.49
[20]	-SMI -DA(Rot, Scal, Mir)	WSI	NDCNN(GoogleNet)		ImageNet	75% / 25%	ILA	97.89	97.64	97.56	97.97
							PLA	97.02	97.23	97.89	97.50
[21]	None	WSI	DR(FV(ConvNet), MNN)	SVM	ImageNet	70% / 30%	ILA	87.7 ± 2.4	87.6 ± 3.9	86.5 ± 2.4	83.9 ± 3.6
							PLA	90.2 ± 3.2	91.2 ± 4.4	87.8 ± 5.3	87.4 ± 7.2
[22]	None	WSI	Integrated		None	70% / 30%	ILA	88.09			
							PLA	88.40			
[23]	None	P(228x228)	NDCNN(AlexNet,trans)		None	Not specified	ILA	82.7%	Not evaluated	Not evaluated	Not evaluated
[24]	None	None	PWT	KNN	None	75% / 25%	ILA	Not evaluated	Not evaluated	Not evaluated	85.62
[25]	RGBT	WSI	JCTF	Linear SVM	None	70% / 30%	PLA	86.88 ± 2.37	88.41 ± 2.73	88.86 ± 3.76	87.55 ± 3.01
[26]	-Res(350x230) -SMI -DA(Rot, Flip)	WSI	NDCNN		None	70% / 30%	ILA	77.5	Not evaluated	Not evaluated	Not evaluated
[27]	-ETB -NDS	WSI	PFTAS	RF	None	Not specified	ILA	81.7 ± 2.8	81.2 ± 2.7	80.7 ± 3.4	81.5 ± 3.1
[28]	- DA(Rot, Mir, Dis)	WSI	Inception v3		ImageNet	70% / 30%	ILA	0.86	Not evaluated	Not evaluated	Not evaluated
[29]	-DA(Zoom, Flip)	Rnd(224x224)	Emax(VGGNet)		ImageNet	80% / 20%	ILA	91.28	91.45	88.57	84.58
[30]	-Res(350x230)	WSI	Eavg(ResNet)		ImageNet	Not specified	PLA	95.0 ± 3.64			
[31]	-Res(224 224) -GSC	WSI	VGG	NN	ImageNet	75% / 25%	ILA	84.0	88.2	87.0	80.3
[32]	None	SQ	BE	SBC	None	70% / 30%	ILA	47.0	40.0	40.0	37.0
[33]	MVD	WSI	DR((Zer, FD, Ent),Rlf)	LSSVM	None	70% / 30%	PLA	87.7	85.8	88.0	84.6
[34]	None	Rnd(64x64)	NDCNN(DenseNet)		Camelyon	75% / 25%	ILA	96.1 ± 3.2	Not considered	Not considered	Not considered
[35]	None	WSI	CT, HI+KM	NDCNN	None	Not specified	ILA	94.40	95.93	97.19	96.00
							F1	95.00	97.00	98.00	96.00
[36]	None	Rnd(64x64)	PFTAS	NPMIL	None	70% / 30%	ILA	87.8 ± 5.6	85.6 ± 4.3	80.8 ± 2.8	82.9 ± 4.1
							PLA	92.1 ± 5.9	89.1 ± 5.2	87.2 ± 4.3	82.7 ± 3.0
[37]	-Res(224x224) - DA(CROP, ANP, FLIP)	WSI	ResNet-152, GoogleNet	SVM	ImageNet	80% / 20%	ILA	81.2 ± 2.5			
[38]	KM	WSI	NDCNN		None	Not specified	ILA	85	90	90	90
							F1	93	93	92	93
[39]	GSC	WSI	HI, KAZE	SVM	None	70% / 30%	ILA	85.9 ± 1.6	80.4 ± 1.4	78.1 ± 2.2	71.1 ± 3.3
							PLA	86.4 ± 2.2	81.6 ± 1.6	77.8 ± 1.6	72.9 ± 2.8
							F1	90.2 ± 1.1	86.5 ± 1.0	84.6 ± 2.2	79.9 ± 3.2
							AUC	94	Not specified	Not specified	Not specified
[40]	SMI	WSI	Inception-V3		ImageNet	70% / 30%	ILA	86.5 ± 3. 7	83.2 ± 2.4	85.4 ± 0.7	80.3 ± 2.2
							PLA	87.6 ± 3.9	82.4 ± 2.7	86.1 ± 0.7	79.7 ± 3.2
							F1	93.0	88.9	89.4	86.9
[41]	KM	WSI	DWT	LSVM	None	Not specified	ILA	93.3			
[42]	-Res(224x224) -DA(Rot, Flip, HS,WS,TR)	WSI	DR(DenseNet-169,XGB)	XGB	ImageNet	70% / 30%	PLA	94.71 ± 0.88	95.9 ± 4.2	96.76 ± 1.09	89.11 ± 0.12
[43]	-GSC -DA	WSI	DR(WSLL-Isomap)	SSAE	None	Not specified	ILA	96.8	98.1	98.2	97.5
[55]	-Res(341x224) - SN - DA(Rot, Flip, WS, HS) -UP	SW(224x224)	ResNet-152		ImageNet	70% / 30%	ILA	98.6	97.9	98.3	97.6
			Emdt(ResNet-152)		ImageNet	70% / 30%	PLA	98.77			
[44]	None	WSI	AlexNet		ImageNet	70% / 30%	ILA	90.69	90.46	90.64	90.96
[45]	Res(227x227)	WSI	AlexNet		ImageNet	Not specified	ILA	90.96 ± 1.59	90.58 ± 1.96	91.37 ± 1.72	91.30 ± 0.74
[46]	None	WSI	DR(PFTAS,PROJ)	QDA	None	70% / 30%	PLA	89.1 ± 2.6	87.3 ± 3.8	88.4 ± 3.6	86.6 ± 2.8
[47]	None	WSI	LPQ	SVM	None	70% / 30%	ILA	91.1	90.7	86.2	84.3
							AUC	0.96	0.96	0.93	0.90
[48]	CE	WSI	Tam	DBN	None	70% / 30%	ILA	88.7	85.3	88.6	88.4
[49]	-Res(370x230) -DA	P(224x224)	NDCNN		None	80% / 20%	ILA	89.52	89.06	88.84	87.67
[50]	None	P(64x64, 32x32)	NDCNN		None	70% / 30%	ILA	82 ± 2.8	86.2 ± 4.6	84.6 ± 3	84 ± 4
							PLA	83 ± 3.2	81 ± 4.2	84.2 ± 3.4	81 ± 2.4
[51]	-Res(224x224) -SMI	WSI	VGG-19	E(SVM)	ImageNet	98% / 2%	ILA	80			
[56]	DA(Rot, Flip)	WSI	Eiter(NDCNN)		None	70% / 30%	ILA	98.33	97.12	97.85	96.15
[52]	SN	WSI	HI	NDCNN	None	85% / 15%	ILA	95.0	96.6	93.50	94.2

Table 5: MSB classification models and results.

5. Magnification-independent binary classification works

After reporting MSB classification works and their results, we will present in this section binary classification models that adopted a magnification-independent approach (MIB). In fact, only two works belongs to this group, and we will present them as follows:

- Section 5.1 presents the first attempt that introduced magnification-independent training to binary classification

- Section 5.2 describes the work that explored the discriminative value contained in each magnification factor subset.

Afterwards, we will summarize in table 6 the best model of each work and its achieved results.

5.1. From a magnification-specific towards a magnification-independent approach

Magnification-independent approach for BreakHis based models was first introduced in [53], where authors proposed to classify histopathological images as either benign or malignant regardless of their magnification factors. To evaluate their MIB reformulation, they elaborated two experiments. In the first one they explored a single task CNN, while in the second one they trained a multi-task CNN. The single task CNN was trained on all magnification subsets combined and tested on each magnification independently to allow its direct comparison with previous magnification-specific works. Results of this comparison proved that this magnification-independent model outperformed some previous magnification-specific works, and more importantly achieved stable results over different magnification. In their second experiment, the authors explored the performance of a multi-task version of the first adopted magnification-independent CNN. This multi-task version was equipped with an additional classifier which serves to predict the magnification level of each input image. Its binary classification results decreased slightly in comparison to those obtained with the single task version. This drop in accuracy was justified by the usefulness of the added magnification factor classifier for the binary classification task.

5.2. Cross-magnification evaluation

Authors in [54] tried to evaluate the discriminative value of each magnification subset independently with a cross-magnification training/test schema. In other words, they tried to explore all possible training/test combinations, where each time a model is trained on a given subset and tested on this same subset or another one. Each model adopted a dual-stage framework. Firstly, it extracted color-texture features. Then, it provided a concatenation of these features to train a majority voting ensemble composed of: SVM , Nearest neighbors, Decision tree and Discriminant Analysis. After evaluating all possible cross-magnification data splittings, these experiments revealed

some interesting findings: Models trained with extreme magnification subsets (×40, ×400) achieved lower classification results with a large variation between different magnification test sets, while those trained with mid-range magnification subsets (×100, ×200) obtained higher results and proved to be more stable over all magnification test sets. The authors justified these facts by the large variability in morphological textures of extreme magnification images (×40, ×400) in comparison to those captured with mid-range magnifications (×100, ×200).

Work	Preprocessing	Patch/Slide	Features extractor	Classifier	Transfer learning	Training/Test	Metrics	Results(%)			
								×40	×100	×200	×400
[53]	-Res(460x460) -DA(Rot, Crop, Flip)	Rnd(100×100)	NDCNN		None	70% / 30%	PLA	83.08 ± 2.08	83.17 ± 3.51	84.63 ± 2.72	82.10 ± 4.42
[54]	None	WSI	JCTF	Emv(classifiers)	None	70% / 30%	PLA	87.2 ± 3.74	88.22 ± 3.28	88.89 ± 2.51	85.82 ± 3.81

Table 6: MIB classification models and results.

6. Magnification-specific multi-category classification works

After reporting all binary classification reformulations, we will present in this section the multi-category classification models trained with a magnification-specific approach (MSM). Mostly, all MSM works were built within a dual-stage framework, where each stage is devoted either to multi-category or binary classification. Thus, these works will be reported according to the logical ordering of their two stages as follows:

- Section 6.1 presents works where the multi-category classification is elaborated before the binary classification
- Section 6.2 describes works where the binary classification is elaborated before the multi-category classification
- Section 6.3 reports works where both classification stages are elaborated independently
- Section 6.4 outlines works where the multi-category classification is elaborated without any binary classification stage

Then, in table 7 we will report the best achieved results in each work.

6.1. Multi-category classification stage before binary classification stage

The first CNN that was designed essentially for multi-category classification is [57]. The latter proposed a two-stages framework composed of a multi-category classification model followed by a binary classification phase. The used CNN was trained only with a multi-category classification output. Then, the binary classification output was intuitively deduced from its corresponding sub-class. Authors of this work noticed that this reformulation has never been explored before due to the high similarities between different sub-classes. According to them, the variance between instances from the same sub-class are greater than the one between those from different sub-classes. To overcome this issue, they integrated a learning constraint to the multi-category CNN. This constraint was added to the output loss function with the aim to control different features similarities during the training, by minimizing the euclidean distance between instances from the same sub-class, while maximizing it between those from different sub-classes or main classes. In addition to the proposed constraint, the authors in this work explored different configurations and also proved the necessity of data augmentation and transfer learning approaches when training a deep learning CNN.

6.2. Binary classification stage before multi-category classification stage

Authors in [55] firstly fine-tuned an ImageNet pre-trained ResNet-152 for the binary classification task. Then, retrained the same fine-tuned ResNet-152 with a multi-category classification output layer. The second stage was composed of two modules, the first one was devoted for benign sub-classes and the second one for malignant sub-classes. After identifying its main class in the binary classification stage, each instance was provided to its corresponding module for sub-class identification. An image was considered correctly classified if only its both stages outputs were correctly classified. For both stages, the decision was made at two levels: At image level; the decision was obtained by merging all extracted patches outputs using majority voting rule. At patient level; the decision was elaborated with a meta-decision tree (MDT)[112] which acts like a trainable ensemble learning approach combining all magnification-specific models. Furthermore, this work revealed that the improvement brought by stain normalization and data augmentation is around 13%.

6.3. Multi-category classification and Binary classification tasks performed independently

In [56], the authors tried to elaborate each one of these two classification tasks independently without any logical link between them. For each classification task, they compared a new-designed CNN to a handcrafted features based model. In the handcrafted features based approach, they evaluated various descriptors encoded with two coding models (bag of words and locality constrained linear coding) in conjunction with an SVM classifier. For the CNN based model, they evaluated a new designed CNN in different use cases: firstly, as a global end-to-end CNN, then as a features extractor or a final classifier provided with prior extracted handcrafted features. Results of this comparison proved the efficiency of the end-to-end CNN model in both classification tasks. Afterwards, they explored the improvement brought by data augmentation as well as ensemble learning with merging prediction outputs captured from the same model at ten different training iterations.

6.4. Multi-category classification without binary classification

Unlike previous works, [58] used a multi-category classification stage only. In fact, it evaluated the performance of three different CNNs in this classification task; namely, a ResNet-v1 model and two Inception variants v1 and v2, all trained in magnification-specific manner. Results of which revealed the ability of the pre-trained ResNet-v1 model to outperform the inception CNNs and achieve a recognition rate of around 95% in such a tedious task. This result was achieved in conjunction with different preprocessing methods including data augmentation and stain normalization, in addition to a fine-tuning strategy where all ResNet layers were retrained on BreakHis multi-category classification task.

Work	Preprocessing	Patch/Slide	Features extractor	Classifier	Transfer learning	Training/Test	Metrics	Results(%)			
								×40	×100	×200	×400
[15]	-Bin	WSI	FD	SVM	None	50 %/ 50 %	F1	96.4 over all magnifications			
[57]	-DAB(IV, Rot, Tr, Flip)	WSI	NDCNN		ImageNet	50% / 50%	ILA	92.8 ± 2.1	93.9 ± 1.9	93.7 ± 2.2	92.9 ± 1.8
							PLA	94.1 ± 2.1	93.2 ± 1.4	94.7 ± 3.6	93.5 ± 2.7
[55]	-Res(341x224) -SN -DA(Rot, Flip, WS, HS) -UP	SW(224x224)	ResNet-152		Im-Break	70% / 30%	ILA	95.6	94.8	95.6	94.6
			Emdt(ResNet-152)		Im-Break	70% / 30%	PLA	96.25 over all magnifications			
[56]	-DA(Rot, Flip)	WSI	Eiter(NDCNN)		None	70% / 30%	ILA	88.23	84.64	83.31	83.98
[58]	-SN(Macenko) -DA(Rot, Flip) -Res(224x224)	WSI	ResNet		ImageNet	80% / 20%	ILA	95.0 over all magnifications			

Table 7: MSM classification models and results.

7. Identifying the best reformulation for BreakHis problem

Identifying the best reformulation for BreakHis dataset is equivalent to find this reformulation from two different standpoints: clinically and practically. Therefore, we will analyse the following aspects:

- Section 7.1 inspects the binary and multi-category classification elements, with the aim to define the best reformulation from a clinical standpoint.

- Section 7.2 analyses the magnification-specific approach in comparison to the magnification-independent one, in order to identify the most adequate reformulation from a practical point of view.

7.1. Binary classification versus multi-category classification

In this part we will present the advantages of adopting a multi-category classification, and the reasons why it has more clinical value than a simple binary classification.

In fact, almost all multi-category classification works [15, 55, 56, 57] include a binary classification task. Instinctively, the encompassing task which is the multi-category classification will always have more clinical value and provides more opportunities to domain experts. Thus, from a clinical standpoint a multi-category classification with an integrated binary classification task is the most adequate reformulation for this problem, and its advantages are the following :

- First, finding the exact malignant sub-category relieves pathologists workloads and guides them to elaborate an optimal therapeutic schedule. This is due to the fact that different treatments exist and determining the exact disease state helps predicting a patient's response to a particular therapy. Notably, lobular tumor (LC) gains a clear benefit from systemic therapy when compared to ductal one (DC) [113].

- Second, the identification of the exact benign lesion sub-type is extremely important because it helps prevent the risk of developing future subsequent breast cancer [114].

7.2. Magnification-independent versus magnification-specific training

From a practical standpoint, to train a multi-category classification model we could either take into consideration the magnification factor of each image (MSM) or train it regardless of this magnification feature (MIM). In this part, we will prove that training this model with magnification-independent approach MIM is in practice more suitable than adopting a magnification-specific approach due to the following factors:

- In a magnification specific approach, one specific model is required for each magnification subset, resulting in more training and adaptation efforts.
- For samples testing in a magnification-specific approach, the magnification factor of each test image must be known a priory to apply the corresponding model. However, this specific information might not be available for all images.
- Unlike a magnification-specific model, a magnification-independent one has the ability to benefit directly from additional training data, and this additional data could be captured with the same or even different magnification factors.
- When authors in some magnification-specific models [30, 55], employed ensembles with their four magnification-specific models, they all achieved higher results. This fact can be interpreted as a reinforcement of the hypothesis claiming that training a model with features from different scales(in magnification-independent manner) can improve its generalization capability.
- A robust system which is intended to be used in a real clinical environment should handle the diversity in microscopic images and not depends on any device setting, because this flexibility is necessary when deploying it in under-developed, developing countries or in rural areas, where microscopes are equipped with limited magnification capabilities.

In fact, most works adopted the MSB reformulation because it is the most straightforward manner to address this problem. However, as we can see from the present discussion, MIM remains clinically and practically the best

reformulation for BreakHis problem. In fact, this reformulation has never been adopted in any work before because of its complexity in learning features from all magnification levels at the same time. In addition, classifying each image among eight different classes with similar patterns is a very tedious task.
In summary, this best MIM reformulation will consist of a multi-category classification model with an implicitly integrated binary classification task, and all trained with a magnification-independent approach.

8. Magnification-independent multi-category classification evaluation with deep learning and learnt lessons

To our knowledge, this work is the first to analyze the magnification-independent multi-category (MIM) classification on BreakHis dataset. In this section, we evaluate this approach experimentally using deep learning and report the learnt lessons as follows:

- Section 8.1 presents the used experimental protocol, deep learning model and adaptation techniques then report the achieved results
- Section 8.2 summarizes the lessons learnt from this experimental study

8.1. Experimental results

In this part we explored the MIM approach using deep learning. In addition, we analyzed the impact brought by fine-tuning different layers of the used CNN, data augmentation and stain normalization as preprocessing techniques. Our MIM system is able to perform a binary classification as well as a multi-category classification. Its a dual-stage classification approach with three different models, one for the binary classification in the first stage, and two other models in second stage, one for benign sub-categories and the other one for malignant sub-categories classification. This formulation was inspired from pathologists workflow that analyzes in the first place whether a given slide is malignant or benign, then depending on its main malignancy, this sample is given to the corresponding sub-category classifier to identify its exact sub-category. To train the three models we used a magnification-independent approach, and for each one we used five trials with different folds, where each fold consists of 70% of BreakHis images as a training set and 30% for test. For each trial, we followed the major constraint adopted

by BreakHis authors in MSB which guarantee that patients who were used for training were not reused during test phase.

Preprocessing: As a pre-processing, we started by downsizing all WSIs from 700×460 into 224×224 to fit in ResNet input. Then, we applied stain normalization to all training images. For stain normalization we explored two different methods: Macenko and Reinhard with a reference image chosen arbitrary. To address the uneven data distribution, we used data augmentation as an upsampling strategy with data balancing constraint between both main classes in the first stage then between four sub-classes in each multi-category model of the second stage. As data augmenters, we used random flipping and random rotation. For the malignant multi-category model we omitted images that belongs to the malignant borderline patient (ID:13412) as it is a label noise sample with two different annotations DC and LC.

Fine-tuning: For each model we used an ImageNet pre-trained ResNet as a base CNN model, and given that the chosen model was pre-trained on a 1000-class classification problem, we replaced the last output layer with a binary output softmax layer for the binary classification model and a four-outputs layer for each one of the two remaining multi-category models. During the training we used WSIs instead of extracted patches. To fine-tune these three CNNs, we adopted a dual-stage fine tuning approach which consists of retraining only the last fully connected layers first, then retraining the whole CNN.

We highlighted each on of the three models and reported their performance in Table 8. Results are presented according to the mean value and standard deviation of classification accuracy at image level (ILA) over five trials. These results presented also the impact of each stain normalization method and the improvement brought by data augmentation as a data balancing solution and the performance of our CNNs at each stage of the adopted fine-tuning strategy.

Model	Fine-tuning stage	Without DA and SN	with DA(Rot,Flip)	with DA(Rot,Flip) and SN(Macenko)	with DA(Rot,Flip) and SN(Reinhard)
Binary	last layers	78.1 ± 2.5	80.3 ± 3.0	75.0 ± 1.3	76.3 ± 0.7
	last layers then all layers	83.5 ± 1.5	88.1 ± 2.1	87.2 ± 1.4	$\mathbf{88.9 \pm 2.5}$
Malignant multi-category	last layers	62.3 ± 1.5	61.2 ± 2.0	60.1 ± 1.2	60.31.3
	last layers then all layers	57.4 ± 2.0	60.1 ± 1.7	56.0 ± 2.3	$\mathbf{63.6 \pm 2.2}$
Benign Multi-category	last layers	33.8 ± 1.9	38.4 ± 2.3	35.0 ± 2.4	37.2 ± 1.5
	last layers then all layers	47.6 ± 1.6	$\mathbf{52.7 \pm 2.3}$	39.0 ± 1.5	41.3 ± 0.9

Table 8: The results of each stage of the proposed MIM model reported independently.

In general, results of the first binary classification stage are very competitive when compared to those reported within the MIB reformulation in

table 6, especially those using the same data partition (70%,30%)[53, 54]. However, the second stage results of the MIM reformulation achieved low accuracy in the malignant sub-classes classification and even lower in the benign sub-classes classification.

8.2. Learned lessons

As it can be seen from this experimental study, the MIM classification accuracy is very low in comparison to the MSM counterpart. This could be explained by the following reasons:

- Learning the differences between the eight subcategories regardless of their magnification levels is much harder for the CNN, especially for the benign sub-categories (minority sub-classes).
- The MSM approach is more adequate to achieve higher results, as it matches better the organization of BreakHis dataset.
- The irregularities generated by the optical microscopic magnifications used to collect BreakHis data. When a ROI in a breast tissue is magnified optically, some new morphological components appears as long as we dive deeper into the tissue. In other words, it is tedious for a deep CNN to learn the most representative features when the same sub-class contains images at different levels with different morphological components.

Given the current BreakHis structure and characteristics, the most reasonable approach is MSM as it maintains the same clinical value as MIM, while it scarifies only its practical counterpart. Otherwise, to achieve good results with the MIM approach itself, several solutions must be explored such as data fusion at the same magnification levels between BreakHis and other available histopathological breast cancer datasets. Since these datasets have different labeling purposes, this data fusion needs more adaptation efforts especially in terms of expert annotation.

9. Conclusion

In this work, we proposed a taxonomy that classifies BreakHis based CAD systems into four reformulations (MSB, MIB, MSM, MIM). Using this taxonomy, we provided a comprehensive survey of all CAD systems that

used BreakHis dataset. We highlighted their main contributions, their used preprocessing methods, their adopted models, their learning strategies in addition to their achieved results at different evaluation levels. To find the best reformulation from a clinical and practical standpoints, we compared these different reformulations and proved that MIM is the best one from both aspects. Then we evaluated this MIM approach using deep learning CNN as it has never been explored before in the literature.

Our study concluded that in the real clinical routine, a CAD system based on the MIM reformulation is the best approach. However, given the current state and organisation of BreakHis dataset, building such an automatic system using CNN does not lead to an acceptable accuracy yet. For instance, we could consider that MSM is the closest reformulation to MIM from a clinical standpoint that fits the actual BreakHis composition. As future works, we suggest that in an MIM based model, more labeling effort would be of high iterest if established on similar datasets such as Bioimaging and MITOSATYPIA with the aim to meet BreakHis structure under a data fusion scenario. Furthermore, we will analyze the potential of other Deep Learning models such as Deep Belief Networks.

10. References

[1] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, F. Bray, Globocan 2012 v1. 0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11. lyon, france: International agency for research on cancer; 2013 (2015).

[2] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012, International journal of cancer 136 (5) (2015) E359–E386.

[3] C. Wang, J. Shi, Q. Zhang, S. Ying, Histopathological image classification with bilinear convolutional neural networks, in: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE, 2017, pp. 4050–4053.

[4] M. Aswathy, M. Jagannath, Detection of breast cancer on digital histopathology images: present status and future possibilities, Informatics in Medicine Unlocked 8 (2017) 74–79.

[5] F. Ghaznavi, A. Evans, A. Madabhushi, M. Feldman, Digital imaging in pathology: whole-slide imaging and beyond, Annual Review of Pathology: Mechanisms of Disease 8 (2013) 331–359.

[6] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, B. Yener, Histopathological image analysis: A review, IEEE reviews in biomedical engineering 2 (2009) 147.

[7] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, Q. Sun, Deep learning for image-based cancer detection and diagnosisa survey, Pattern Recognition.

[8] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, A. Campilho, Classification of breast cancer histology images using convolutional neural networks, PloS one 12 (6) (2017) e0177544.

[9] M. Veta, P. J. Van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. Larsen, J. S. Vestergaard, A. B. Dahl, et al., Assessment of algorithms for mitosis detection in breast cancer histopathology images, Medical image analysis 20 (1) (2015) 237–248.

[10] J. Xu, L. Xiang, R. Hang, J. Wu, Stacked sparse autoencoder (ssae) based framework for nuclei patch classification on breast cancer histopathology, in: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), IEEE, 2014, pp. 999–1002.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING 63 (7) (2016) 1455.

[12] M. Veta, J. P. Pluim, P. J. Van Diest, M. A. Viergever, Breast cancer histopathology image analysis: A review, IEEE Transactions on Biomedical Engineering 61 (5) (2014) 1400–1411.

[13] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, in: Australasian joint conference on artificial intelligence, Springer, 2006, pp. 1015–1021.

[14] F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in: Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE, 2016, pp. 2560–2567.

[15] A. Chan, J. A. Tuszynski, Automatic prediction of tumour malignancy in breast cancer with fractal dimension, Royal Society open science 3 (12) (2016) 160558.

[16] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, L. Heutte, Deep features for breast cancer histopathological image classification, in: Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on, IEEE, 2017, pp. 1868–1873.

[17] Y. Song, J. J. Zou, H. Chang, W. Cai, Adapting fisher vectors for histopathology image classification, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE, 2017, pp. 600–603.

[18] M. A. Kahya, W. Al-Hayani, Z. Y. Algamal, Classification of breast cancer histopathology images based on adaptive sparse support vector machine, Journal of Applied Mathematics and Bioinformatics 7 (1) (2017) 49.

[19] K. Das, S. P. K. Karri, A. G. Roy, J. Chatterjee, D. Sheet, Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE, 2017, pp. 1024–1027.

[20] B. Wei, Z. Han, X. He, Y. Yin, Deep learning model based breast cancer histopathological image classification, in: Cloud Computing and Big Data Analysis (ICCCBDA), 2017 IEEE 2nd International Conference on, IEEE, 2017, pp. 348–353.

[21] Y. Song, H. Chang, H. Huang, W. Cai, Supervised intra-embedding of fisher vectors for histopathology image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 99–106.

[22] V. Gupta, A. Bhavsar, An integrated multi-scale model for breast cancer histopathological image classification with joint colour-texture features, in: International Conference on Computer Analysis of Images and Patterns, Springer, 2017, pp. 354–366.

[23] S. Akbar, M. Peikari, S. Salama, S. Nofech-Mozes, A. Martel, Transitioning between convolutional and fully connected layers in neural networks, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 143–150.

[24] A. A. Samah, M. F. A. Fauzi, S. Mansor, Classification of benign and malignant tumors in histopathology images, in: Signal and Image Processing Applications (ICSIPA), 2017 IEEE International Conference on, IEEE, 2017, pp. 102–106.

[25] V. Gupta, A. Singh, K. Sharma, A. Bhavsar, Automated classification for breast cancer histopathology images: Is stain normalization important?, in: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, Springer, 2017, pp. 160–169.

[26] E. M. Nejad, L. S. Affendey, R. B. Latip, I. Bin Ishak, Classification of histopathology images of breast into benign and malignant using a single-layer convolutional neural network, in: Proceedings of the International Conference on Imaging, Signal Processing and Communication, ACM, 2017, pp. 50–53.

[27] M. Sharma, R. Singh, M. Bhattacharya, Classification of breast tumors as benign and malignant using textural feature descriptor, in: Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on, IEEE, 2017, pp. 1110–1113.

[28] J. Chang, J. Yu, T. Han, H.-j. Chang, E. Park, A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer, in: e-Health Networking, Applications and Services (Healthcom), 2017 IEEE 19th International Conference on, IEEE, 2017, pp. 1–4.

[29] W. Zhi, H. W. F. Yueng, Z. Chen, S. M. Zandavi, Z. Lu, Y. Y. Chung, Using transfer learning with convolutional neural networks to diagnose

breast cancer from histopathological images, in: International Conference on Neural Information Processing, Springer, 2017, pp. 669–676.

[30] J. Sun, A. Binder, Comparison of deep learning architectures for h&e histopathology images, in: Big Data and Analytics (ICBDA), 2017 IEEE Conference on, IEEE, 2017, pp. 43–48.

[31] S. Cascianelli, R. Bello-Cerezo, F. Bianconi, M. L. Fravolini, M. Belal, B. Palumbo, J. N. Kather, Dimensionality reduction strategies for cnn-based classification of histopathological images, in: International Conference on Intelligent Interactive Multimedia Systems and Services, Springer, 2017, pp. 21–30.

[32] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, Y. Zhao, Size-scalable content-based histopathological image retrieval from database that consists of wsis, IEEE journal of biomedical and health informatics 22 (4) (2018) 1278–1287.

[33] S. Chattoraj, K. Vishwakarma, Classification of histopathological breast cancer images using iterative vmd aided zernike moments & textural signatures, arXiv preprint arXiv:1801.04880.

[34] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation equivariant cnns for digital pathology, arXiv preprint arXiv:1806.03962.

[35] A.-A. Nahid, Y. Kong, Histopathological breast-image classification using local and frequency domains by convolutional neural network, Information 9 (1) (2018) 19.

[36] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, P. Honeine, Multiple instance learning for histopathological breast cancer image classification, Expert Systems with Applications 117 (2019) 103–111.

[37] G. Zhang, M. Xiao, Y.-h. Huang, Histopathological image recognition with domain knowledge based deep features, in: International Conference on Intelligent Computing, Springer, 2018, pp. 349–359.

[38] A.-A. Nahid, M. A. Mehrabi, Y. Kong, Histopathological breast cancer image classification by deep neural network techniques guided by local clustering, BioMed research international 2018.

[39] D. Sanchez-Morillo, J. González, M. García-Rojo, J. Ortega, Classification of breast cancer histopathological images using kaze features, in: International Conference on Bioinformatics and Biomedical Engineering, Springer, 2018, pp. 276–286.

[40] Y. Benhammou, S. Tabik, B. Achchab, F. Herrera, A first study exploring the performance of the state-of-the art cnn model in the problem of breast cancer, in: Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, ACM, 2018, p. 47.

[41] R. Karthiga, K. Narasimhan, Automated diagnosis of breast cancer using wavelet based entropy features, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2018, pp. 274–279.

[42] V. Gupta, A. Bhavsar, Sequential modeling of deep features for breast cancer histopathological image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2254–2261.

[43] S. Pratiher, S. Chattoraj, Manifold learning & stacked sparse autoencoder for robust breast cancer classification from histopathological images, arXiv preprint arXiv:1806.06876.

[44] B. Du, Q. Qi, H. Zheng, Y. Huang, X. Ding, Breast cancer histopathological image classification via deep active learning and confidence boosting, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 109–116.

[45] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, Ü. Budak, Transfer learning based histopathologic image classification for breast cancer detection, Health information science and systems 6 (1) (2018) 18.

[46] P. Alirezazadeh, B. Hejrati, A. Monsef-Esfehani, A. Fathi, Representation learning-based unsupervised domain adaptation for classification of breast cancer histopathology images, Biocybernetics and Biomedical Engineering.

[47] J. A. Badejo, E. Adetiba, A. Akinrinmade, M. B. Akanle, Medical image classification with hand-designed or machine-designed texture

descriptors: A performance evaluation, in: International Conference on Bioinformatics and Biomedical Engineering, Springer, 2018, pp. 266–275.

[48] A.-A. Nahid, A. Mikaelian, Y. Kong, Histopathological breast-image classification with restricted boltzmann machine along with backpropagation., Biomedical Research 29 (10) (2018) 2068–2077.

[49] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, D. Sheet, Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 578–581.

[50] K. Kumar, A. C. S. Rao, Breast cancer classification of image using convolutional neural network, in: 2018 4th International Conference on Recent Advances in Information Technology (RAIT), IEEE, 2018, pp. 1–6.

[51] E. M. Nejad, L. S. Affendey, R. B. Latip, I. B. Ishak, R. Banaeeyan, Transferred semantic scores for scalable retrieval of histopathological breast cancer images, International Journal of Multimedia Information Retrieval 7 (4) (2018) 241–249.

[52] A.-A. Nahid, Y. Kong, Histopathological breast-image classification using concatenated r–g–b histogram information, Annals of Data Science (2018) 1–17.

[53] N. Bayramoglu, J. Kannala, J. Heikkilä, Deep learning for magnification independent breast cancer histopathology image classification, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 2440–2445.

[54] V. Gupta, A. Bhavsar, Breast cancer histopathological image classification: is magnification important?, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[55] Z. Gandomkar, P. C. Brennan, C. Mello-Thoms, Mudern: Multi-category classification of breast histopathological image using deep residual networks, Artificial intelligence in medicine.

[56] D. Bardou, K. Zhang, S. M. Ahmad, Classification of breast cancer based on histology images using convolutional neural networks, IEEE Access 6 (2018) 24680–24693.

[57] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, S. Li, Breast cancer multi-classification from histopathological images with structured deep learning model, Scientific reports 7 (1) (2017) 4172.

[58] M. A. Nawaz, A. A. Sewissy, T. H. A. Soliman, Automated classification of breast cancer histology images using deep learning based convolutional neural networks, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY 18 (4) (2018) 152–160.

[59] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on pattern analysis and machine intelligence 24 (7) (2002) 971–987.

[60] Z. Guo, L. Zhang, D. Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Transactions on Image Processing 19 (6) (2010) 1657–1663.

[61] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: International conference on image and signal processing, Springer, 2008, pp. 236–243.

[62] R. M. Haralick, K. Shanmugam, et al., Textural features for image classification, IEEE Transactions on systems, man, and cybernetics (6) (1973) 610–621.

[63] L. P. Coelho, A. Ahmed, A. Arnold, J. Kangas, A.-S. Sheikh, E. P. Xing, W. W. Cohen, R. F. Murphy, Structured literature image finder: extracting information from text and images in biomedical literature, in: Linking Literature, Information, and Knowledge for Biology, Springer, 2010, pp. 23–32.

[64] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: An efficient alternative to sift or surf, in: Computer Vision (ICCV), 2011 IEEE international conference on, IEEE, 2011, pp. 2564–2571.

[65] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in neural information processing systems, 2006, pp. 1473–1480.

[66] A. Tharwat, Linear vs. quadratic discriminant analysis classifier: a tutorial, International Journal of Applied Pattern Recognition 3 (2) (2016) 145–180.

[67] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, ACM, 1992, pp. 144–152.

[68] V. Lepetit, P. Fua, Keypoint recognition using randomized trees, IEEE transactions on pattern analysis and machine intelligence 28 (9) (2006) 1465–1479.

[69] B. B. Mandelbrot, The fractal geometry of nature, Vol. 1, WH freeman New York, 1982.

[70] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE transactions on signal processing 62 (3) (2014) 531–544.

[71] K. Kira, L. A. Rendell, A practical approach to feature selection, in: Machine Learning Proceedings 1992, Elsevier, 1992, pp. 249–256.

[72] J. Zhu, S. Rosset, R. Tibshirani, T. J. Hastie, 1-norm support vector machines, in: Advances in neural information processing systems, 2004, pp. 49–56.

[73] C. Liao, S. Li, Z. Luo, Gene selection using wilcoxon rank sum test and support vector machine for cancer classification, in: International Conference on Computational and Information Science, Springer, 2006, pp. 57–66.

[74] P. F. Alcantarilla, A. Bartoli, A. J. Davison, Kaze features, in: European Conference on Computer Vision, Springer, 2012, pp. 214–227.

[75] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, 1990, pp. 396–404.

[76] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[77] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[78] M. N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Transactions on image processing 14 (12) (2005) 2091–2106.

[79] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[80] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[81] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International conference on machine learning, 2014, pp. 647–655.

[82] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[83] M. Ringnér, What is principal component analysis?, Nature biotechnology 26 (3) (2008) 303.

[84] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 245–250.

[85] K. Michalak, H. Kwasnicka, Correlation based feature selection method, International Journal of Bio-Inspired Computation 2 (5) (2010) 319–332.

[86] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks., in: CVPR, Vol. 1, 2017, p. 3.

[87] T. Chen, C. Guestrin, Xgboost: Reliable large-scale tree boosting system. arxiv 2016; 1–6, DOI: http://dx. doi. org/10.1145/2939672.2939785.

[88] S.-J. Lee, T. Chen, L. Yu, C.-H. Lai, Image classification based on the boost convolutional neural network, IEEE Access 6 (2018) 12755–12768.

[89] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European conference on computer vision, Springer, 2010, pp. 143–156.

[90] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 3828–3836.

[91] Z. Gao, L. Wang, L. Zhou, J. Zhang, Hep-2 cell image classification with deep convolutional neural networks, IEEE journal of biomedical and health informatics 21 (2) (2017) 416–428.

[92] Y. Song, H. Chang, Y. Gao, S. Liu, D. Zhang, J. Yao, W. Chrzanowski, W. Cai, Feature learning with component selective encoding for histopathology image classification, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 257–260.

[93] Y. Song, Q. Li, H. Huang, D. Feng, M. Chen, W. Cai, Low dimensional representation of fisher vectors for microscopy image classification, IEEE transactions on medical imaging 36 (8) (2017) 1636–1649.

[94] T. Cohen, M. Welling, Group equivariant convolutional networks, in: International conference on machine learning, 2016, pp. 2990–2999.

[95] M. Sun, T. X. Han, M.-C. Liu, A. Khodayari-Rostamabad, Multiple instance learning convolutional neural networks for object recognition, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 3270–3275.

[96] R. Venkatesan, P. Chandakkar, B. Li, Simpler non-parametric methods provide as good or better results to multiple-instance learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2605–2613.

[97] N. Zeng, Z. Wang, H. Zhang, W. Liu, F. E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognitive Computation 8 (4) (2016) 684–692.

[98] H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, IEEE Transactions on Systems, man, and cybernetics 8 (6) (1978) 460–473.

[99] J. B. Tenenbaum, V. De Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, science 290 (5500) (2000) 2319–2323.

[100] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A. M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649.

[101] C. Senaras, M. N. Gurcan, Deep learning for medical image analysis, Journal of pathology informatics 9.

[102] M. Bakator, D. Radosav, Deep learning and medical diagnosis: A review of literature, Multimodal Technologies and Interaction 2 (3) (2018) 47.

[103] M. Biswas, V. Kuppili, L. Saba, D. Edla, H. Suri, E. Cuadrado-Godia, J. Laird, R. Marinhoe, J. Sanches, A. Nicolaides, et al., State-of-the-art review on deep learning in medical imaging., Frontiers in bioscience (Landmark edition) 24 (2019) 392–426.

[104] A. Maier, C. Syben, T. Lasser, C. Riess, A gentle introduction to deep learning in medical image processing, Zeitschrift für Medizinische Physik 29 (2) (2019) 86–101.

[105] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annual review of biomedical engineering 19 (2017) 221–248.

[106] M. H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: Achievements and challenges, Journal of digital imaging (2019) 1–15.

[107] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson, H. Liu, Deep learning and alternative learning strategies for retrospective real-world clinical data, NPJ digital medicine 2 (1) (2019) 43.

[108] G. Murtaza, L. Shuib, A. W. A. Wahab, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfiqar, G. Raza, N. A. Azmi, Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges, Artificial Intelligence Review (2019) 1–66.

[109] G.-S. Fu, Y. Levin-Schwartz, Q.-H. Lin, D. Zhang, Machine learning for medical imaging, Journal of healthcare engineering 2019.

[110] M. A. Mazurowski, M. Buda, A. Saha, M. R. Bashir, Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri, Journal of Magnetic Resonance Imaging 49 (4) (2019) 939–954.

[111] M. J. Shensa, The discrete wavelet transform: wedding the a trous and mallat algorithms, IEEE Transactions on signal processing 40 (10) (1992) 2464–2482.

[112] L. Todorovski, S. Džeroski, Combining classifiers with meta decision trees, Machine learning 50 (3) (2003) 223–249.

[113] R. Barroso-Sousa, O. Metzger-Filho, Differences between invasive lobular and invasive ductal carcinoma of the breast: results and therapeutic implications, Therapeutic advances in medical oncology 8 (4) (2016) 261–266.

[114] M. Guray, A. A. Sahin, Benign breast diseases: classification, diagnosis, and management, The oncologist 11 (5) (2006) 435–449.

Appendix A. Abbreviations dictionary

Preprocessing:	
Res(hxw)	Resizing original images to a new size (hxw)
SMI	Subtracting mean images
SN	Stain normalization
SN(x)	Stain normalization using method x
UP	Minority class upsamling
RGBT	RGB channels information transformation
NDS	Nuclei Detection and Segmentation using region growing technique
ETB	E_AHE and TB_HAT techniques
GSC	Grey-scale conversion
MVD	Multilevel variational mode decomposition
KM	K-mean clustering
CE	Contrast Enhancement
JPEG	Conversion to JPEG
CD	Color distortion
HP	Hue permutation
BS	Brightness saturation
Bin	Binarization
DA(x)	Data augmentation using x
DAB(x)	Data augmentation with a data balancing purpose using x
IV	Intensity variation
Rot	Rotation
Tr	Translation
Flip	Flipping
Scal	Scaling
Mir	Mirroring
Dis	Distortion
Zoom	Zooming
CROP	Cropping
ANP	Adding noisy points
HS	Height shift
WS	Width shift
Res	Resizing
Patch/Slide extraction:	
WSI	Whole slide Image
P(axa)	Patches extraction of size (axa)
Rnd(axa)	Random patches extraction of size (axa)
SW(axa)	Sliding window patches extraction of size (axa)
MKV(N)	Extraction of N patches using MKV framework
SQ	Division into non-overlapping square tiles
Features extractors:	
PFTAS	Parameter-Free Threshold Adjacency Statistics
FD	Fractal dimension
Ent	Entropy
PWT	Pyramid-Structure Wavelet Transform
JCTF	Joint color-texture features
Zer	Zernaike moments
CT	Contourlet Transform
HI	Histogram information
KAZE	KAZE features
DWT	Discrete wavelet transform
LPQ	Local Phase Quantization
Tam	Tamura features
BE	Binarization encoding
FV(x)	Fisher vector encoded using a model x
DR(f,x)	Dimentionality reduction of features f using a method x
WRS	Wilcoxon rank sum
MNN	Multilayer neural network
Rlf	Relief
PROJ	Projection into an invariant space
Classifiers:	
QDA	Quadratic Linear Analysis
SVM	Support Vector Machine
LR	Logistic Regression
Integrated	An integrated model composed of the features extractors and classifiers that achieved the best results at each magnification level
ASSVM	Adaptive Sparse Support Vector Machine
KNN	k-nearest neighbours
RF	Rotation Forest
NN	Nearest-neighbour
LSSVM	Least squares support vector machines
NPMIL	Non-parametric Multi-instance learning
LSVM	linear SVM
XGB	XGBoost classifier
SSAE	Stacked sparse autoencoders
DBN	Deep Belief Network
SBC	Similarity based comparison using Hamming distance
NDCNN	A new designed CNN model
NDCNN(x)	A new designed CNN model inspired from x
NDCNN(x,trans)	A new designed CNN model constituted of x with an integrated transition layer
Ensemble learning:	
E(C)	Ensemble of classifiers C
Emv(C)	Ensemble of classifiers C using majority voting rule
Eavg(C)	Ensemble of classifiers C using average rule
Esum(C)	Ensemble of classifiers C using sum rule
Emax(C)	Ensemble of classifiers C using max rule
Emdt(C)	Ensemble of classifiers C using MDT
Transfer learning:	
ImageNet	The used model was pre-trained on ImageNet
Camelyon	The model was pre-trained on Camelyon16 dataset images
Im-Break	The used CNN that was fine-tuned on the BreakHis binary classification task after ImageNet, is retrained once again on the BreakHis multi-category classification task
Metrics:	
ILA	Image Level Accuracy
PLA	Patient Level Accuracy
F1	F1-score
AUC	Area Under Curve