Artificial Intelligence in Medicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

MuDeRN: Multi-category classification of breast histopathological image using deep residual networks

Ziba Gandomkar[a,*], Patrick C. Brennan[a], Claudia Mello-Thoms[a,b]

[a] Image Optimisation and Perception, Discipline of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia
[b] Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 28 September 2017
Received in revised form 20 February 2018
Accepted 13 April 2018

Keywords:
Benign breast lesion
Breast cancer
Breast cancer subtypes
Deep learning
Deep residual networks

ABSTRACT

Motivation: Identifying carcinoma subtype can help to select appropriate treatment options and determining the subtype of benign lesions can be beneficial to estimate the patients' risk of developing cancer in the future. Pathologists' assessment of lesion subtypes is considered as the gold standard, however, sometimes strong disagreements among pathologists for distinction among lesion subtypes have been previously reported in the literature.
Objective: To propose a framework for classifying hematoxylin-eosin stained breast digital slides either as benign or cancer, and then categorizing cancer and benign cases into four different subtypes each.
Materials and methods: We used data from a publicly available database (BreakHis) of 81 patients where each patient had images at four magnification factors (×40, ×100, ×200, and ×400) available, for a total of 7786 images. The proposed framework, called MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) consisted of two stages. In the first stage, for each magnification factor, a deep residual network (ResNet) with 152 layers has been trained for classifying patches from the images as benign or malignant. In the next stage, the images classified as malignant were subdivided into four cancer subcategories and those categorized as benign were classified into four subtypes. Finally, the diagnosis for each patient was made by combining outputs of ResNets' processed images in different magnification factors using a meta-decision tree.
Results: For the malignant/benign classification of images, MuDeRN's first stage achieved correct classification rates (CCR) of 98.52%, 97.90%, 98.33%, and 97.66% in ×40, ×100, ×200, and ×400 magnification factors respectively. For eight-class categorization of images based on the output of MuDeRN's both stages, CCRs in four magnification factors were 95.40%, 94.90%, 95.70%, and 94.60%. Finally, for making patient-level diagnosis, MuDeRN achieved a CCR of 96.25% for eight-class categorization.
Conclusions: MuDeRN can be helpful in the categorization of breast lesions.

1. Introduction

Breast cancer (BCa) is the most common non-skin cancer among women worldwide. Despite the increase in the incidence rate of BCa over last few decades, the mortality rate from BCa in the developed countries has been decreased due to improvements in treatment options and early detection through screening mammography [1]. For every 1000 women who have participated in screening mammography, 15.6 to 17.5 need a needle biopsy [2] but only one in four are diagnosed with BCa [3]. Therefore, each year, pathologists evaluate a large number of breast histopathological slides, from which only about 25% contains malignancy, and benign lesions are far more prevalent.

The diagnoses made by pathologists on the cases are usually considered as the gold standard for further treatment of the patients. However, recent studies have shown that the pathologists might disagree with an expert consensus-derived reference diagnoses in distinguishing benign cases from cancer [4–6]. In [5], 6900 individual case diagnoses made by 115 pathologists were compared with an expert consensus-derived ground truth and 17% of benign cases with atypia and 3% of benign cases without atypia were misdiagnosed as ductal carcinoma in situ or invasive carcinoma, while 10% of invasive carcinoma or ductal carcinoma in situ were misdiagnosed as benign cases with or without atypia. Also, it was shown that pathologists who inter-

* Corresponding author. Postal address: Room 726, Level 7, Brain and Mind Centre, 94 Mallett Street, Camperdown, NSW, 2050, Australia.
E-mail address: ziba.gandomkar@sydney.edu.au (Z. Gandomkar).

https://doi.org/10.1016/j.artmed.2018.04.005
0933-3657/

pret a smaller number of cases per week and those working as general pathologists make more diagnostic errors than experts [3–5]. Allison et al. [4] divided underlying reasons for disagreement among the pathologists into three categories, which were pathologist-related, diagnostic coding/study methodology-related, and specimen-related. Among pathologist-related factors, "professional differences of opinion on features meeting diagnostic criteria" was ranked first. Computer-assisted analysis can be helpful in reducing the discrepancies in benign/malignant classification by providing an objective classification.

Recently, with the advent of whole slide imaging and production of digital histopathology slides, many researchers have started developing computer-aided detection tools for classification of breast slides as benign or malignant [7]. For example, Weyn et al. [8] used wavelet-based, Haralick, intensity-based, and morphological features extracted from segmented nuclei and their surrounding for classification of breast histopathological slides as benign or malignant, and achieved a correct classification rate (CCR) of 79% for case-based classification. In [9], 84 features (morphological, intensity-based, and textural) extracted from isolated nuclei were utilized for classifying images as either benign or malignant. A sensitivity of 97% and a specificity of 94% has been achieved. However, both methods are computationally expensive as the epithelial nuclei were segmented first. Unlike these methods, Yang et al. [10] extracted textural features using a texton-based approach without segmenting the structures in slides. Using this method, 89% of images were classified correctly.

Pathologists are responsible for not only identifying whether a lesion is malignant or benign but also determining the benign or cancer subtypes, as both benign and malignant breast lesions encompass different subcategories. Different treatment options are available for BCa patients and determining the BCa subtype could be helpful in predicting the patient's response to therapy; for example, invasive lobular cancer gains a clear benefit from systemic therapy when compared to invasive ductal cancer [11]. The correct recognition of benign lesion type is also important because the patient's risk of developing subsequent BCa varies among different types of benign lesions [12].

Cserni et al. [13] showed that there are discrepancies among pathologists for determining benign lesion subtypes. The study asked six pathologists to classify benign lesions into three categories, namely fibroadenoma, phyllodes tumor, and anything other these subtypes. The overall Cohen's kappa for categorizing was 0.48, which suggests a moderate agreement [14]. Lawton et al. [15] investigated the agreement among ten pathologists for distinguishing fibroadenomas from phyllodes tumors and found that there was 100% agreement only in 53% of cases. In [16], the inter-observer agreement for classification of invasive breast carcinoma was studied and the highest agreement rates among 13 pathologists were achieved for mucinous, lobular, and tubular subtypes, with agreement rates of 96.0%, 78.7%, and 78.0% respectively.

Similar to the malignant/benign classification, computer-assisted analysis could help pathologists to increase diagnostic agreement in multi-class categorization of lesions. Despite the importance of determining the lesion subtype, only a few previous studies aimed at automatic classification of breast lesions into different subtypes. In [10], six subtypes of BCa were divided into two subgroups: cancer class I, which contains ductal carcinoma in situ and lobular carcinoma in situ, and cancer class II, containing invasive ductal carcinoma, invasive lobular carcinoma, lymph-node-negative metastasis, and soft tissue metastasis. Using a texton-based approach, images were classified into three classes, i.e. benign, cancer type I, and cancer II and a CCR of 80% was achieved.

Recently there has been growing research on the application of Deep Learning (DL) in medical image segmentation and classification [17]. DL revolutionized machine learning in the past few years. Conventionally, in supervised machine learning techniques, the discriminant features were selected based on domain-specific knowledge and computer algorithms determined the optimal decision boundary in the feature space. However, in DL, computers learn the optimal feature representations from the data. Although deep neural networks originated from previously existing artificial neural networks[18], training the deep architectures have recently become practical due to emergence of high-performance GPU computing, which makes it feasible to train networks with many hidden layers in a reasonable time. AlexNet, one the earliest convolution neural networks, won the ImageNet challenge in December 2012 [19]. Since then, further progress in the architecture of networks and learning algorithms has been made [17]. Similar to other machine learning fields, DL algorithms have been widely used for analyzing medical images, especially in segmentation tasks such as anatomical structures segmentation in retinal images [20],or organ segmentation in abdominal CT images [21,22]. DL has also been utilized for classification tasks such as classification of breast lesions as benign or malignant on mammograms [23] or tomosynthesis [24]. In musculoskeletal image analysis, DL was adopted it for different application such as age assessment on x-ray and MRI images or vertebrae localization and identification on CT and MRI images. The promising results are also achieved in other areas such as brain [25] or cardiac [26] images.

In the field of digital pathology, DL has been used for nuclei detection, segmentation, and classification [27]. For example, Xu and Huang [28] utilized a distributed deep neural network architecture to detect cells in whole-slide high-resolution histopathological images. In [29], a multiscale convolutional network was used for accurate segmentation of cervical cytoplasm and nuclei. DL has been also adopted for organ segmentation in large histopathological images such as segmentation of colon glands [30] or neuronal structures [31]. Furthermore, it has been utilized for classification purposes; for example, colon cancer classification [32], thyroid cytopathology classification [33], or Gleason grading of prostate cancer images [34].

DL algorithms were also used for analyzing breast histopathology slides. In [35], it was used to detect mitotic figures within breast slides. In the tutorial on DL provided in [27], AlexNet network schema was used to address the segmentation of nuclei, epithelium, and tubule as well as detection of invasive ductal carcinomas, lymphocyte, and mitosis. Spanhol et al. [36] used AlexNet for classifying breast histopathological images as benign or malignant. In [34], breast cancer areas were detected by incorporating shearlet features inside a convolutional neural networks. Cruz-Roa et al. [37] used a DL approach for automatic detection of invasive ductal carcinoma tissue. In [38], a context-aware stacked convolutional neural network architecture was used for classifying whole slide images as benign, ductal carcinoma in situ, or invasive ductal carcinoma.

This paper focuses on three tasks which are: (i) classification of breast histopathological images as benign or malignant, (ii) categorization of malignant images as ductal carcinoma, lobular carcinoma, mucinous carcinoma, or papillary carcinoma; and (iii) classification of benign images as adenosis, fibroadenoma, phyllodes tumour, or tubular adenoma. Previously Han et al. [39] used GoogLeNet [40] for classifying breast histopathological images into similar eight categories and used majority voting for patient classification. Although this aimed at addressing an almost similar problem, we improved both the image-level (i.e. considering each image individually without incorporating the patient information for decision making) and the patient-level (i.e. appointing a single label to each patient by aggregating the class labels assigned to all images of that patient) classification CCRs. This was achieved by first carrying out stain normalization as a pre-processing step. Secondly, we used a deeper network and a two-stage classifier and

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

thirdly, we utilized a meta-decision tree (MDT) [41] for making the patient-level diagnosis based on the four magnification factors.

In this study, we propose a framework, called MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) for classifying patients based on hematoxylin-eosin stained breast digital slides either as benign or cancer, and then categorizing cancer and benign cases into four different subtypes each. MuDeRN uses a very deep residual neural network [42], i.e. a deep residual network with 152 layers (ResNet-152), for classification of breast histopathological images as benign or malignant. Images were acquired in four different magnification factors and for each factor, a separate network has been trained. Malignant images were then subdivided into four subcategories while benign images were classified as four benign subtypes. Eventually, the final diagnosis for a patient was made by combining outputs of networks for different magnification factors using an MDT [41]. It considers the confidence level of the label given by the networks for each magnification factor, and also the CCR of the assigned labels to select the best magnification factor for making a patient-level diagnosis. The major contributions of this study are (i) using ResNet for the first time for differentiation of benign and malignant subtypes in addition to multi-class classification of breast histopathological slides; (ii) proposing a novel framework for combining outputs based on different magnification factors to make the ultimate diagnosis for a patient using a trainable approach (i.e. MDT); and (iii) investigating whether the stain normalization step can be replaced by color or contrast augmentation. Previous studies used non-trainable methods, such as majority voting, for aggregating the image-level diagnoses to produce patient-level diagnosis. We hypothesized that a MDT as a trainable approach outperforms majority voting, which is the most common non-trainable approach.

2. Materials and methods

In this section, the dataset we used and MuDeRN's steps are discussed. This study was exempt from the requirement for approval by the Human Research Ethics Committees at the University of Sydney because the data was obtained from a publicly available dataset and all images were de-identified.

The steps of MuDeRN for categorization of lesion subtypes are shown in Fig. 1. Briefly, for each patient a set of images from four magnification factors (x40, x100, x200, and x400) were available. To mitigate the color variations, images were normalized by two dif-

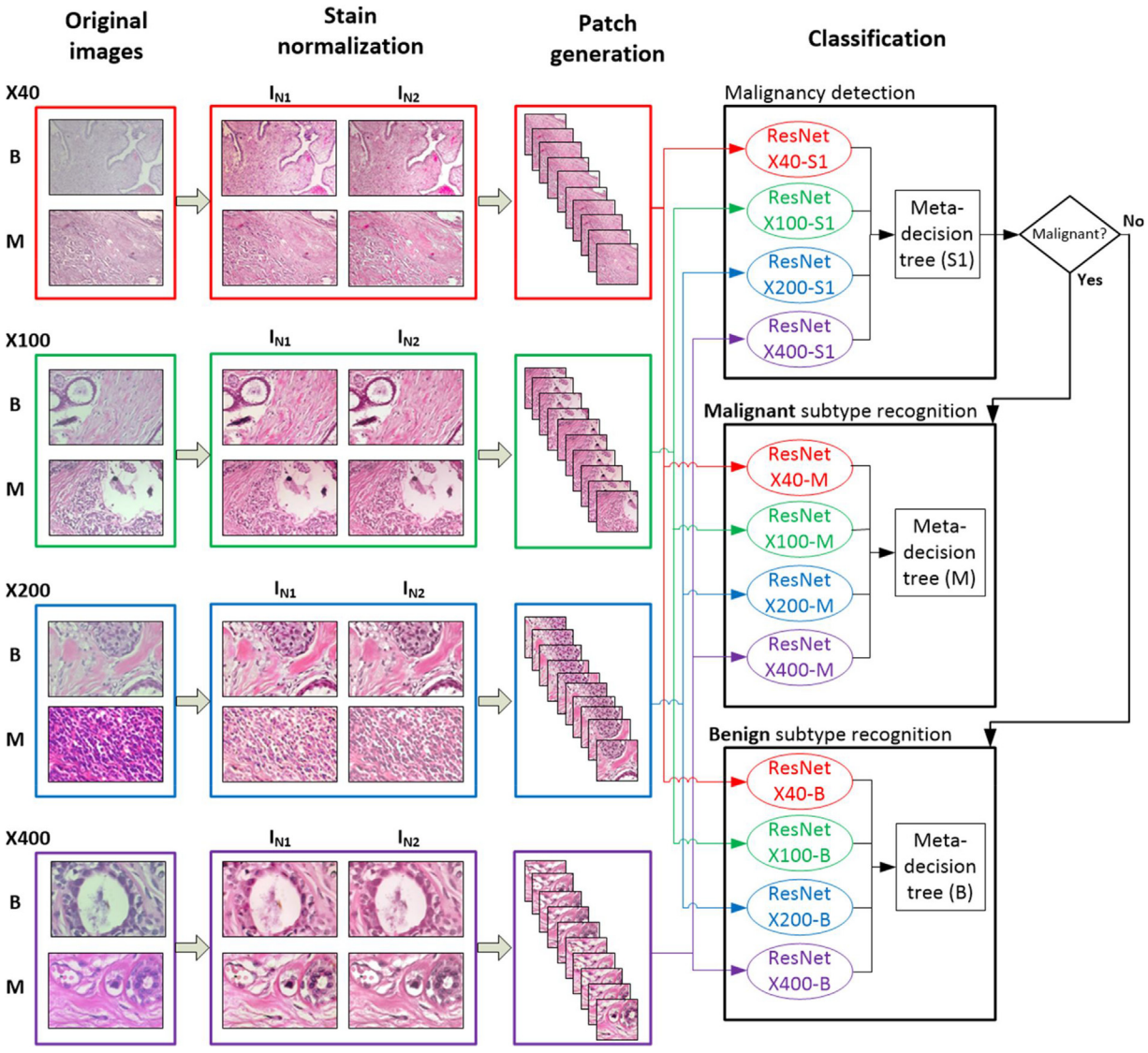

Fig. 1. The steps of MuDeRN.

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

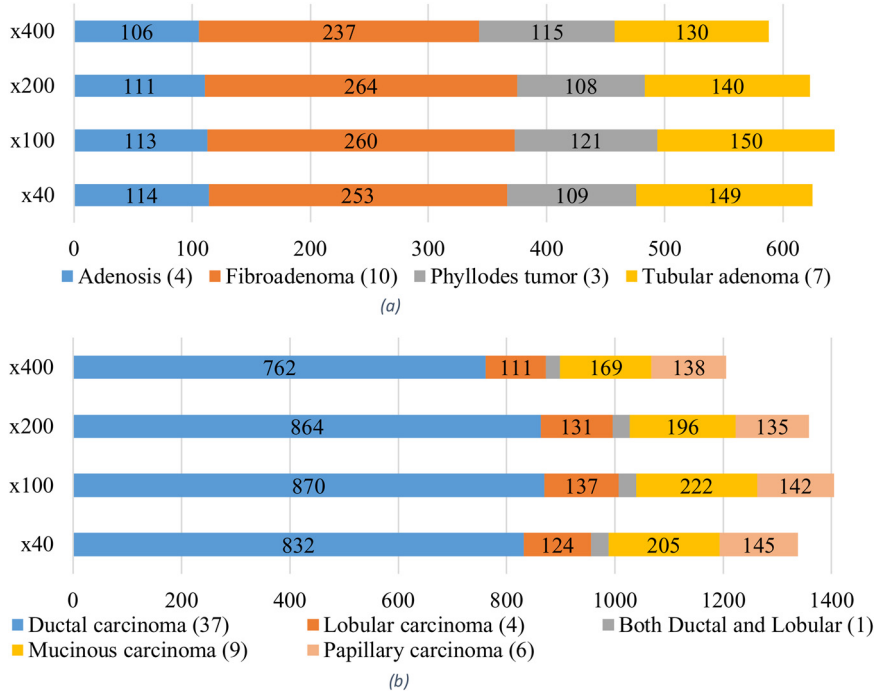

Fig. 2. Distribution of (a) benign (b) malignant images by magnification factor and class, number of patients in each category is shown in parentheses. Numbers in each row represent number of images in BreakHis for all patients per each subtype.

ferent methods, which are explained in 2–3. From each normalized image, square image patches were extracted and fed into a ResNet for classification. For each magnification factor, a separated network was trained. The classification was done in two stages. In the first stage (S1) patches were classified either as benign or malignant and an image-level decision was made by using weighted majority voting. On average 24 images were available per patient per each magnification factor. For making patient-level diagnosis, an MDT [41] was used to combine the probabilities of malignancy given to the different images of a patient. The second stage consists of two modules, M and B. The images which were classified as malignant by the first stage were fed into module M where they were subdivided into four cancer subtypes, while those classified as benign were inputted to module B, where they classified into four categories. The architectures of the module in the second stage was almost identical to that of the first stage but for the four classes.

2.1. Dataset

To evaluate the performance of MuDeRN, we used the BreakHis database [43], which is a publicly available dataset of hematoxylin-eosin (HE) stained breast histopathological slides. The images were acquired in four visual magnification factors, namely ×40, ×100, ×200, and ×400 with the effective pixel size of 0.49 μm, 0.20 μm, 0.10 μm, and 0.05 μm respectively. The images were stored in a format of three-channel red–green–blue (RGB) TrueColor (24-bit color depth, 8 bits per color channel) color space. For each patient, a pathologist identified a few diagnostically-relevant regions of interest (ROI). The undesired areas, such as text annotations or black border, were removed and the images were cropped to a dimension of 700 × 460 pixels. Finally, out-of-focus images were also discarded.

On average 24.23, 25.28, 24.46, and 22.15 images were available per patient in ×40, ×100, ×200, and ×400 respectively. The BreakHis database is comprised of 82 folders corresponding to 82 patients, however, one of the patients (Patient ID: 13,412) was a borderline case (has features of both ductal and lobular carcinoma) and hence was placed in both ductal and lobular groups. This patient was included in benign/malignant classification but excluded for tumor sub-type recognition. Fig. 2 shows the distribution of images over different sub-types.

2.2. Deep residual network

Deep neural networks are cascades of layers of nonlinear processing units that form a hierarchy corresponding to multiple levels of data representation, starting by learning low-level features (such as edges and lines) to higher-level features (which combine the low-level features with elements of tissue). They are increasingly popular models for automatic classification and segmentation in medical image analysis when large-scale label data is available.

The AlexNet [19] is one of the earliest deep neural networks which contains five convolutional layers followed by fully connected layers. The AlexNet aimed at classification of images into 1000 object categories. Unlike the conventional neural networks which use hyperbolic tangent as the activation function, AlexNet uses rectified linear units (ReLU) as they are several times faster.

Recent evidence indicated that deeper networks (a network with more layers), such as GoogLeNet (22 layers) [40] achieved bet-

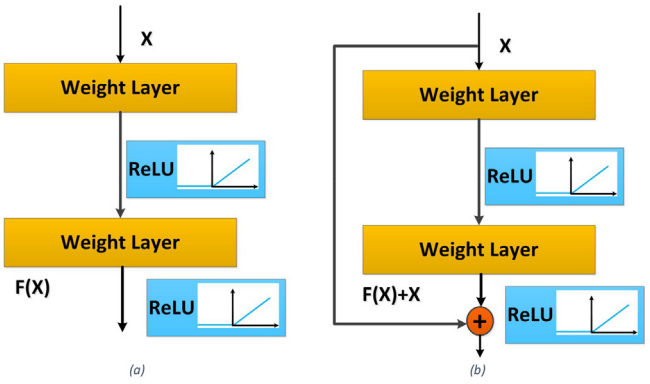

Fig. 3. Building block of (a) a plain net (b) a ResNet. ReLU is a rectified linear unit.

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

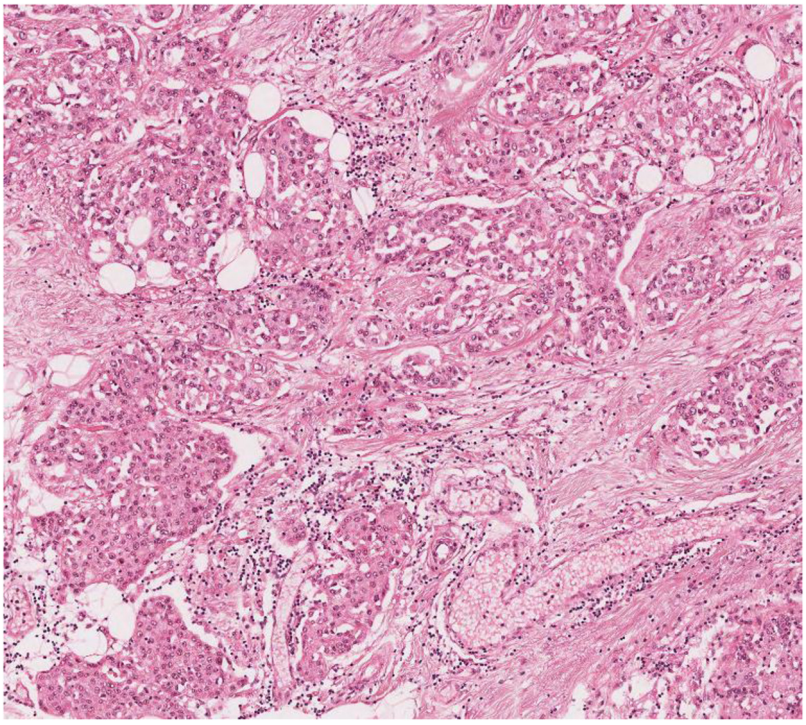

Fig. 4. This target image was used as a reference image to which all the images were mapped.

ter results on the ImageNet dataset. However, simply stacking more convolutional layers will not lead to a lower classification error and "overly deep plain networks" have higher training error compared to their shallower counterpart [42]. This phenomenon, which is called the degradation problem, could be due to the optimization difficulty of finding the weights of all hidden layers in a feasible time when the network is overly deep. To tackle this problem, ResNet was proposed in [42]. In Fig. 3, the building block of the ResNet is compared to the plain network. In the plain network (Fig. 3(a)) the mapping from input to output can be represented by the nonlinear H(x) function. Assume that instead of H(x), F(x) = H(x)-x is used. As shown in Fig. 3(b), at the output of the second weight layer x was added to the F(x) and then their sum passes to the ReLU. He et al. [42] showed that adding this shortcut from input to the output of the stacked layers could tackle the optimization difficulties of the deeper networks as the gradient can flow directly from later layers to the earlier layers. Because in ResNet the shortcut connections simply perform identity mapping, extra parameters (and hence extra computational complexity) is not added to the optimization task.

ResNet models with 50, 101, and 152 layers were trained and tested on the ImageNet dataset [42]. As expected the error was the least for the ResNet with 152 layers. Previously ResNet-152 was used for analysis of histopathological slides in [44]. It achieved an overall accuracy of 93.0% for classification of colorectal whole-slide images into six classes (five types of colorectal polyps and the normal class) and performed better than ResNet with 50 and 101 layers. Therefore, in this study, we used the ResNet-152 for classification.

2.3. Stain normalization

Inconsistencies in color are major issues in analysis of histopathological slides. The inconsistencies could be due to different reasons such as the use of different chemicals for staining, variations in color concentrations, or differences in scanners from different vendors. Different algorithms have been suggested for stain normalization. Each algorithm has its own advantages and limitations and works better for a group of images but has some flaws when applied to other images. Hence, here we used two different stain normalization methods and produced two stain normalized images, I_{N1} and I_{N2}, for each image.

I_{N1} was produced using a stain normalization algorithm based on histogram specification [45], where the images from a patient were transformed to a set of new images so that the histograms of the output images in different color channels approximately match the target image histogram in the corresponding channel. Fig. 4 indicates the target image that was used in this study. This image was selected from the Mitosis-Atypia database[1] based on the opinion of a pathologist who was asked to select an image with an appropriate staining which includes lumen, stroma and epithelial cells and does not have any artifact or tissue folding. We did the histogram matching for the stack of images of a given patient in each magnification factor rather than doing the normalization image by image. This was done to mitigate the visual artifacts in the images due to the assumption of the histogram specification that the proportion of pixels in each color is almost identical in the source and target image. If each time only a single image had been considered, this assumption might have been violated as only a very small tissue area with limited tissue elements and hence limited number of colors would have been taken into account. The second approach used for stain normalization was first proposed in [46], where the mean and standard deviation of each channel of the images were matched to that of the reference image by using a set of linear transformation in La*b* color space.

2.4. Stage 1: benign/malignant classification

As shown in Fig. 1, in the first stage binary classification was performed for detecting the malignant and benign cases. In this section, first we explain the steps of MuDeRN for classifying images

[1] http://mitos-atypia-14.grand-challenge.org/.

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

of a patient as malignant or benign. Then we explain in detail how MuDeRN was trained and tested by using BreakHis database.

The first stage was composed of four ResNets, where each was trained for classifying breast histopathological images of a specific magnification factor. The input size of the ResNets was 224×224 while the images in BreakHis database were 700×460 pixels. Therefore, both I_{N1} and I_{N2}, were resized to 341×224. Then five overlapping patches with size of 224×224 were extracted from each one of the stained normalized images by using a sliding window. Thus, for each image, ten patches (five from I_{N1} and five from I_{N2}) were extracted. For each image in each magnification factor, the probability that the image is belonging to the jth class, was found using (1).

$$S_j = \frac{\sum_{p=1}^{10} s_{j,p}}{10}, \; j \in \{M, B\}, \; p \in \{1, .., 10\} \tag{1}$$

M and B represent malignant and benign classes and $s_{j,p}$ shows the probability of the pth patch of the image belonging to the jth class. The label with the highest probability was assigned to the image. Therefore, the label assigned to image was $J = \arg\max_j S_j$. The image-level CCR was calculated as the number of images in each magnification factor which were classified correctly by the total number of images in that magnification factor.

In order to make the final diagnosis for a patient, image-level diagnoses for four magnification factors have been combined using a MDT. As stated earlier, on average we have approximately 24 images per patient per magnification factor. For each magnification level with N images, the probability that a case is belonging to the jth class, was found using (2).

$$CL_j = \frac{\sum_{i=1}^{N} S_{j,i}}{10}, \; j \in \{M, B\}, \; p \in \{1, .., 10\} \tag{2}$$

In (2), $S_{j,i}$ represents the probability that ith image in that specific magnification level is belonging to the jth class. We calculated CL_j as suggested in (2) for all magnification levels. Assume CL_j^x indicates the CL_j when calculated for the xth magnification factor.

In each magnification factor, the class maximizing CL_j^x, $J_x = \arg\max_j CL_j^x$, was found and for each patient, we had $\left\{J_{x40}, \max_j CL_j^{x40}, J_{x100}, \max_j CL_j^{x100}, J_{x200}, \max_j CL_j^{x200}, J_{x400}, \max_j CL_j^{x400}\right\}$. This was then fed into the MDT (S1) to make the final diagnosis for a patient. We used MDTs as suggested in [41] for combining multiple classifiers. The MDT classified the patient based on the confidence level of the label assigned in different magnification factors and the discriminative power of each magnification factor when used for classifying the validation set. So, the MDT specifies the best magnification factor for a specific test patient. For example, if the confidence of the ResNet in the magnification factor of ×200 was 100% (the highest compared to all other magnification factors) for a test data and the CCR of the ResNet in ×200 for the validation set was also 100%, then the MDT would assign the label of test data as outputted by the ResNet in the magnification factor of ×200.

For evaluation of MuDeRN, 27-fold cross validation was used. The main reason that we used cross-validation instead of splitting the data into training, validation, and test sets, was that there was not enough data available for training the ResNets and Metadata without losing significant testing capability.

Eighty-one patients were randomly divided into 27 subsets, from which 24 contained one benign patient and two cancer patients and 3 contained three cancer patients. Also, we made sure that all subsets contained at least one patient with ductal carcinoma. The list of patients in each fold is provided as an excel file in the supplementary materials. This was done because for some categories we had only a few patients and we wanted to make sure that all these patients were not grouped into one subset. Each time one of the sets served as the test set and the rest of the patients were split into the training set with 70 patients and the validation set with 8 patients. The parameters of the ResNets were estimated based on the training data while the validation data was used for training the MDT for the patient-level diagnosis.

As the number of benign images were approximately half of the malignant images, we upsampled the benign class by extracting twice as many patches from the training and validation sets. Therefore, for each benign image, 20 patches were extracted from I_{N1} and I_{N2} while ten patches were extracted from each malignant image.

Training the ResNet from scratch (i.e. the random initialization of the network's weight and training the model) requires a very large scale dataset as the network has a large number of parameters. Therefore, we fine-tuned the ResNet, previously trained on the ImageNet (1.2 M label images) by continuing to train it on the training set using stochastic gradient descent (SGD) with back-propagation with a small learning rate (0.0001) for 50 epochs. Although the classification task done on the ImageNet data set is completely different to breast histopathological image classification, due to lack of training data, making a model from scratch was not feasible. In [39], it was shown that the accuracy of GoogLeNet for eight-class classification of breast histopathological slides was higher for fine-tuning in comparison with training from scratch. Before start training, the last classification layer of the pre-trained ResNet model was removed and replaced with a classification layer with only two classes as ResNet had been trained for classifying image into 1000 categories.

Image augmentation artificially creates training images through different ways of processing or combination of multiple processing and is usually required to improve the performance of deep networks and avoid overfitting of the network to the training set. Here the training data was augmented by random combination of image rotation by 90°, 180°, or 270°, flipping about horizontal or vertical axes, and random horizontal and vertical shifting between ± 10 pixels. In each epoch, one pass over the training patches was completed and in each pass, the image patches were randomly augmented.

The validation set has only eight members, which makes training the MDT difficult. Therefore, an up-sampling strategy was required. Assume $P_j | j \in \{M, B\}$ indicates total number of patches extracted from an image. Here we set P_M and P_B to 10 and 20 respectively as number of benign images were approximately half of that of the malignant images. For upsampling, we randomly grouped patches from different images of a single patient together in a way that each group contains only one patch from a particular image; hence for each patient, P_j samples were generated. For example, when the validation set contains 2 benign and 6 malignant patients, in total 100 samples, i.e. 2 (number of benign patients) ×20 (P_B) + 6 (number of malignant patients) × 10 (P_M). Therefore, the confidence level of each group of patches was found by averaging the score given to all patches that included in that class. Similarly to (2), for xth magnification factor and pth set of patches, the class with the maximum value, $\hat{J}_{x,p} = \underset{j}{argmax} \hat{CL}_{j,p}^x$, was found. Therefore we have P_j samples for each patient and the data used for training the MDT had the format of $\left\{\hat{J}_{x,p}, \max_j \hat{CL}_{j,p}^x | x \in \{x40, x100, x200, x400\}, \; j \in \{M, B\}, \; p \in \{1, \ldots, P_j\}\right\}$.

After estimating the parameters of the ResNets in four magnification factors and also the parameters of the MDT, for each test image, ten patches were randomly selected from I_{N1} and I_{N2}. Patches in each magnification factor were

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

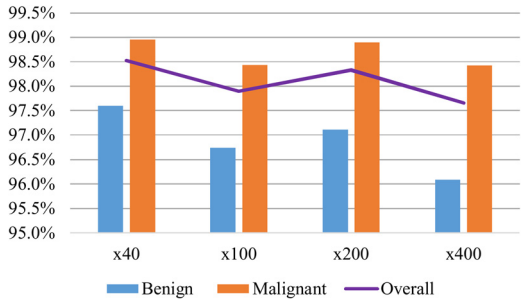

Fig. 5. Accuracy of ResNets in the first stage for malignant/benign classification of images in different magnification factors.

fed into the corresponding ResNet, which outputted $cl_{j,p}^{i,x}$. Using (1) and (2), the value of CL_j^x was calculated for each image. Finally for each patient in the test subset $\left\{J_{x40}, \underset{j}{maxCL_j^{x40}}, J_{x100}, \underset{j}{maxCL_j^{x100}}, J_{x200}, \underset{j}{maxCL_j^{x200}}, J_{x400}, \underset{j}{maxCL_j^{x400}}\right\}$ was generated and inputted to the trained MDT to make the final diagnosis for each patient.

2.5. Stage 2: differentiation of lesion sub-types

Based on the decision made in the first stage, malignant images were fed into the module M as shown in Fig. 1, and categorized into four cancer subtypes, while benign images were inputted to the module B and classified into four categories.

The architecture of both modules were almost identical to what was used in the first stage. However, there were two differences between them. First, in S1 we used the model pre-trained on the ImageNet data set as a starting point for the fine-tuning. Here, we used the ResNet trained at the S1 as the starting point. We hypothesized that the first few layers of the ResNet already learned the low-level features for describing the breast histopathological images and hence it could be a better starting point.

Second, the number of classes in this stage is four per each module, therefore we have $j \in \{1, 2, 3, 4\}$. In the module for processing benign images, the total number of patches extracted from images in the training and validation sets for each class, P_j, was 24 for adenosis and phyllodes tumor, 10 for fibroadenoma, and 16 for tubular adenoma. Similarly, the minority classes were upsampled in the module M, resulting to 10, 70, 40, and 60 patches per image for ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma respectively. Each time half of the patches were extracted from I_{N1} while the other half were extracted from I_{N2}. In S1, a sliding window was used to extract overlapping image patches; here we used a sliding window but we also randomly rotated (0°, 90°, or 180°) and flipped (none, horizontal, or vertical) the image patches as well. This was done because for the minority classes we extracted 20–35 patches per normalized image and different patches are almost identical with a very slight shift.

3. Results

3.1. Image-level performance

The values of CCRs for the ResNets processing images of different magnification factors in the first stage are shown in Fig. 5. For all magnification factors, CCRs for the benign category were lower than those for the malignant one. The differences between CCR of benign and malignant categories are significant for all magnification factors (×40: $z = -2.32$, p-value = 0.020; ×200: $z = -2.48$, $p = 0.013$; ×200: $z = -2.88$, p-value = 0.004; ×400: $z = -3.07$, p-value = 0.002). The overall CCR varied across different magnification factors and ranged from 97.9% to 98.3%, but the differences among overall CCRs at different magnification factors were not statistically significant.

Table 1
Accuracy of the MuDeRN's module for the recognition of different benign subtypes in various magnification factors. The highest CCR for each class is shown in bold.

	×40	×100	×200	×400
Adenosis	89.47%	87.61%	**89.19%**	86.79%
Fibroadenoma	98.02%	96.15%	**99.24%**	97.47%
Phyllodes tumor	**94.50%**	92.56%	93.52%	94.78%
Tubular adenoma	94.63%	**96.67%**	95.00%	95.38%
Overall	95.04%	94.10%	**95.51%**	94.56%

Table 2
Accuracy of the MuDeRN's module for the recognition of different malignant subtypes in various magnification factors. The highest CCR for each class is shown in bold.

	×40	×100	×200	×400
Ductal carcinoma	**98.44%**	97.82%	98.26%	98.29%
Lobular carcinoma	97.58%	97.81%	97.71%	**98.20%**
Mucinous carcinoma	96.59%	96.85%	**97.45%**	96.45%
Papillary carcinoma	95.86%	**96.48%**	96.30%	96.38%
Overall	97.78%	97.52%	**97.89%**	97.80%

Table 3
Accuracy of MuDeRN in different magnification factors. The highest CCR for each class is shown in bold.

	×40	×100	×200	×400
Adenosis	82.46%	85.84%	**89.19%**	86.79%
Fibroadenoma	96.44%	92.31%	**95.83%**	93.25%
Phyllodes tumor	**93.58%**	93.39%	92.59%	89.57%
Tubular adenoma	**93.29%**	92.67%	91.43%	92.31%
Ductal carcinoma	**97.96%**	97.13%	97.57%	97.38%
Lobular carcinoma	95.16%	95.62%	**96.18%**	94.59%
Mucinous carcinoma	**95.61%**	94.59%	95.41%	93.49%
Papillary carcinoma	95.17%	96.48%	95.56%	**95.65%**
Overall	95.60%	94.89%	**95.69%**	94.63%

The CCR values for the recognition of different benign and malignant subtypes in various magnification factors are listed in Tables 1 and 2 respectively. In Table 1, we included all benign images, no matter whether they were detected correctly by the malignant/benign classification module. Similarly, all malignant images regardless of their labels from the first stage were included in Table 2. Table 3 indicates the overall CCRs when outputs from both stages were combined. Hence, the image was considered correctly classified when it was assigned the appropriate label by the first stage and then in the next stage, the subtype was correctly identified. Here we presented results for both four-class categorization (Tables 1 and 2) and eight-class categorization (Table 3) separately because we wanted to show the performances of stand-alone modules for distinction among benign subtypes and distinguishing among cancer subtypes, as some pathologists might prefer to classify images into benign and malignant themselves and use MuDeRN to aid in the distinction among subtypes, so that the error from the first stage does not propagate in the classification done by the second stage.

As shown in Table 1, overall CCR value of ×200 magnification factor was the highest, however, the differences among the CCR values for different magnification factors were not significant. For adenosis and Fibroadenoma, the ×200 magnification factor achieved the highest CCR value while for the Phyllodes tumor class, the highest CCRs was achieved when images from the lowest magnification factor were classified. For tubular adenoma CCR value of ×100 magnification factor was the highest.

As shown in Table 2, for the malignant subtypes the overall CCR value for ×200 magnification factor was the highest, which is similar to the results presented in Table 1 for the recognition of different

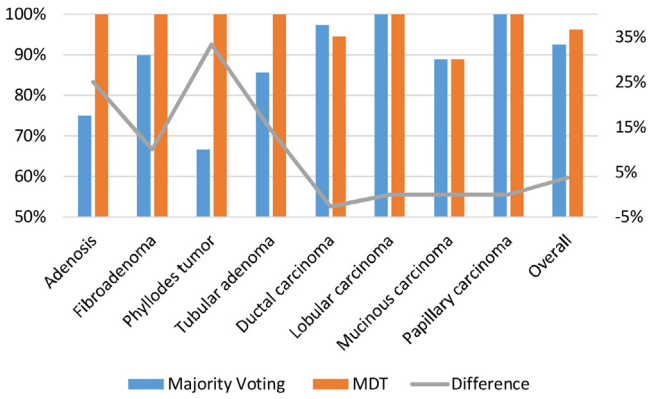

Fig. 6. Comparison of adopting the meta-decision tree for making patient-level diagnosis with majority voting.

benign subtypes. Here for each class, a different magnification factor resulted in the highest CCR for each class.

Table 3 shows the MuDeRN's CCR at the image-level in each magnification factor and for each class. As shown, for two benign subtypes and two cancer subtypes, ×100 magnification factor performed the best and for the rest of the subtypes, ×200 magnification factor outperformed the others.

3.2. Patient-level performance

Finally, as explained in Section 2.5, MDTs were used to make a patient-level diagnosis through combining the outputs from different magnification factors. A CCR of 98.77% was achieved in the first stage for classification of patients either as benign or malignant while the overall CCR for the patient-level diagnosis was 96.25%. Training the MDT added an extra computational burden to the algorithm, so one might question the advantage of MDT over a non-trainable aggregation strategy. The most common non-trainable way to aggregate image-level classification and produce patient-level diagnose is majority voting of the image-level results. We compared the performance of MDT with that of majority voting to explore the added benefit of using MDT for aggregating the image-level results. Fig. 6 shows the comparison of two aggregation methods. As shown overall the MDT method performed about 4% better than non-trainable method. As it can be seen, the differences between two methods varied for different diseases. The advantage of MDT was more prominent for different types of benign diseases.

3.3. Impact of preprocessing

The main preprocessing stages of MuDeRN is stain normalization and image augmentation. To explore the benefit of normalizing stain, we trained and tested MuDeRN without the stain normalization. A heuristic alternative approach to the stain normalization could be augmenting the images regarding color and/or contrast. Similar to the stain normalization, the main purpose of this type of augmentation is providing robustness against changes in color or brightness. Therefore, we explored the impact of different augmenters and their combinations. The first considered augmenter we applied was random contrast adjustment with a factor between 0.5 and 1.5, which was sampled from a uniform distribution. The contrast augmenter can process images either channel-wise or be applied identically to all three channels. We investigated both these strategies. We also investigated the impact of color augmentation by adding noise (adding a randomly selected value between −0.1 and 0.1) to the hue and saturation channel of the HSV representation. The third augmenter randomly multiplied each image by a randomly selected value between 0.75 and 1.25. This was leading to darker or brighter images. The last augmenter added a randomly selected value between −25 to 25 to channels of the image. We also considered the combination of all these augmenters.

Between two stain normalization techniques that we used, the first one was more common for normalizing the histopathological images. Therefore, we also evaluated the performance of ResNets by using all augmenters when the images were normalized using the first approach and the second set of normalized images were not included. Originally we used rotation, flipping, and translation for image augmentation. Here, we also investigated the added benefit of color and contrast augmenters (all augmenters previously described) to the MuDeRN performance (both stain normalized images were considered). Fig. 7 shows the values of CCR for different classes and magnification factors when different pre-processing strategies were adopted.

As shown, the effect of different data augmenters varied across different magnification level and different classes. When averaged across all classes and magnification levels, the percentage drop in CCR, i.e. $(CCR_{New}\text{-}CCR_{MuDeRN})/CCR_{MuDeRN}$, was the highest (13.01%) when the stain normalization step was omitted while no extra data augmenters were adopted. Using the multiplying and adding augmenters, decreased this percentage drop from 13.01% to 10.94% and 9.36%, respectively. Adopting contrast augmenters reduced the percentage drop slightly more than these two augmenters. The average percentage drop was 8.40% when the contrast augmenter was applied identically to different channels and 8.90% when it was applied to each channel differently. As shown, for almost all classes and magnification factors the most effective augmenter for replacing the color normalization was the color augmenter. Its corresponding average percentage drop was only 4.08%. Finally, when all augmenters were adopted and the stain normalization step was eliminated, the average percentage drop across all classes and magnification levels was only 3.06%. Among different classes, the changes in CCRs as a result of adopting a different preprocessing strategy were more prominent for phyllodes tumor and lobular carcinoma.

As indicated in Fig. 7, when contrast and color data augmenters were also utilized, the performance of MuDeRN increased only 0.43% compared to its original performance. This was consistently observed for different magnification levels and classes. The results also suggested that when all augmenters were adopted, the CCRs did not change notably no matter whether both or only one of the normalized images were used.

4. Discussion

In this paper, MuDeRN was proposed to classify patients based on HE stained breast histopathological images either as benign or malignant, and also categorize them into eight classes, representing different subtypes of benign lesions and carcinomas. MuDeRN consisted of two stages. The first stage had a single module composed of four ResNets, where each one dealt with a specific magnification factor and a MDT for combining image-level predictions to classify patients either as benign or malignant. The second stage was comprised of two modules, one for categorizing malignant images into four subtypes and one for classifying benign images into four subcategories. MuDeRN was tested on a database containing 7786 images in four magnification factors from 81 patients. It achieved an average CCR of 98.10% over all magnification factors for classifying the images as benign or malignant while an average CCR of 95.15% for classifying images into eight classes. At the patient-level, MuDeRN achieved a CCR of 98.77% for malignant/benign classification, and 96.25% for the eight-category classification.

As shown in Table 3, the CCR values varied among different subtypes. This could be due to the fact that numbers of patients in

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

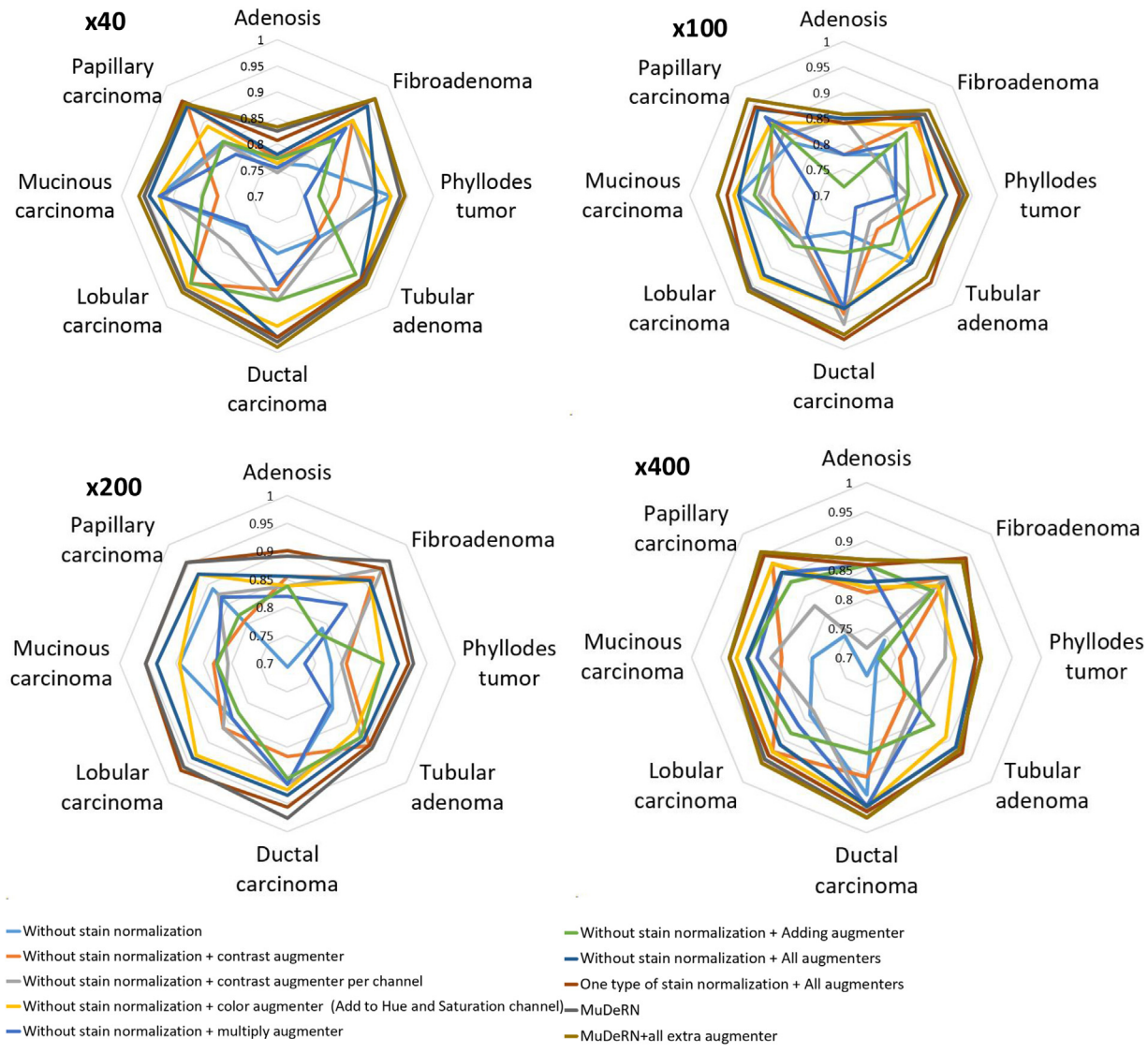

Fig. 7. The impact of different pre-processing strategy on values of CCR among different classes and magnification levels.

different subtypes were not similar. For example, ductal carcinoma which achieved the highest CCR, had the highest number of patients as well. By providing larger number of cases, the ResNets learn the characteristics of lesion better. For the adenosis subtype, the CCR was the lowest. That could be because adenosis has different subtypes (i.e. sclerosing, tubular, apocrine, microglandular) and larger number of cases is required so that the network learns features of different variations of the disease.

For binary classification, the ResNet processing the images from the lowest magnification factor, i.e. × 40, achieved the highest overall CCR. This is in line with the results obtained in [43] where the conventional classifiers and the textural features were used for the binary classification the BreakHis database. The pathologists also start by evaluating the slide in the lowest magnification factor and then zoom in to a few areas at the higher magnification factors for making the final diagnosis. This behavior could explain the fact that the images in the lowest magnification factor of the database (i.e. ×40 magnification factor) had slightly more discriminative power compared to other magnification factors. For eight-category classification, the ResNet analyzing images of ×200 magnification factor achieved the highest CCR value. This could imply that the ×40 magnification factor is more informative for making decisions about the existence of malignancy, however further information, especially cytological features, from higher magnification factors is required for identification of lesion subtypes.

We investigated the impact of the stain normalization on the performance of the proposed framework. Our results showed that lack of the stain normalization preprocessing led to drop in the CCR, which was more considerable in the highest magnification level (×400). For training MuDeRN, images were augmented involved using random rotation, flipping, and translation. We also investigated whether further image augmentation by random color and contrast transformation could replace the stain normalization preprocessing and also improve MuDeRN performance. The results showed that different types of augmentation were able to mitigate the drop in CCR as a result of eliminating the stain normalization step. The improvement in CCR as a result of color augmentation was the highest. Also, the performance of MuDeRN did not increase considerably compared to its original performance, when further augmentations were adopted. However, the benefits of these augmentation strategies should be further investigated when histopathological images were collected from multiple centers with more variations in staining protocols.

A few recent studies used AlexNet and GoogleNet for binary-class classification of images from the same database [36,39,43,47]. In [43], a conventional machine learning pipeline where hand-

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005

crafted features were extracted was used and an average CCR of 82.23% across all magnifications was achieved. In [36,39,47], average CCR values ranging between 83.25% and 84.85% were obtained once AlexNet was adopted. Han et al. [39] used GoogLeNet and achieved an average CCR of 96.08% across all magnifications. These values suggest that MuDeRN improved the CCRs compared to the previous studies. However, the results are not fully comparable, as the validation protocol of our study is different from those studies.

The only previous study that dealt with multi-class classification of images was presented in [39].The main differences of MuDeRN from the framework presented in [39] is using different architecture of neural networks, involving the stain normalization preprocessing, using different image resizing strategy, and adopting a trainable approach for combining the results of different classifiers. In this study, we used ResNet, which is deeper than the network adopted in [39] (i.e. GoogLeNet and AlexNet). Also, the images were downsized to 256×256 in [39]. As the aspect ratio of the original images in BreakHis was about 1.52, resizing them could change the aspect ratio of the structures within the tissue and result in altering some informative features of the image. Here we extracted square patches from the images and then used the weighted majority voting for image-level classification. Finally, we used MDT, which is a trainable approach, for aggregating the outputs of classifiers from different magnification factors.

It was observed that in cases exhibiting some mixed features of two or more diseases, the classifier in one particular magnification factor correctly recognizes the true class while in the other magnification levels the classifiers might misrecognize the case. Usually in the process of differential diagnosis, pathologists make a few candidate diagnoses for a case. They utilize evidence in a particular magnification level, where the differences among candidate diagnoses are more obvious, to rule out a candidate diagnosis. For example, phyllodes tumor is characterized by its leaf-like architecture compared to fibroadenoma. To a human observer, the difference between these two benign subtypes is more apparent in $\times 40$ magnification factor. As they are both benign fibro-epithelial tumors, the appearance of epithelial cells in $\times 200$ and $\times 400$ shares some similarities. In the high magnification factors, mildly increased stromal cellularity is suggestive of phyllodes tumors as compared with a fibroadenoma. However, in some cases stromal overgrowth might not be present. Even in these cases, the presence of elongated, branching and cleft-like ducts and a leaf-like architecture, which can be seen more obviously in $\times 40$ magnification factor, provides a clue for supporting the diagnosis of phyllodes tumor rather than fibroadenoma. In concordance with this, in our study ResNets misclassified some images of phyllodes tumor as fibroadenoma in $\times 200$ and $\times 400$ while the images of the same patients were classified correctly in x40. We hypothesized that the MDT mimics pathologists' step-by-step decision making process. We compared the performance of a common non-trainable, i.e. majority voting, with that of the MDT (trainable model). The overall accuracy of the MDT was approximately 4% better than the non-trainable model and the added benefit of using MDT varied among different classes. By including in situ cases to the database, which are more challenging, the differences between the performances of the MDT and non-trainable method could become even more significant.

This study has a number of limitations. First, cases with non-invasive BCa (ductal carcinoma in situ and lobular carcinoma in situ) were not included in the BreakHis database. These types of BCa are pre-invasive and demonstrate features between benign and invasive cancer, making diagnoses of these cases more difficult. Although the results obtained in this study are promising, that could be to some extent because of lack of the borderline cases in the database. Therefore, including in situ cases could be a possible avenue for future work. Secondly, the regions of interest were manually selected by the pathologists in the BreakHis database, which makes MuDeRN semi-automatic. Therefore, one potential future work could be adding a preprocessing stage which automatically selects the diagnostically relevant areas of the whole slide images. Also, the target image for stain normalization was selected manually based on opinion of a pathologist. Selecting a different image as the target image could affect the appearance of stain normalized images and some variability exists in this selection, which will propagate through the framework. Thirdly, in the BreakHis database only four benign subtypes and four cancer subtypes were considered, however, both benign lesions and invasive cancer have other subtypes which should be included. In addition, for some subtypes, only a few cases were included and the performance of MuDeRN should be investigated on a larger database including more patients from theses subtypes. Also, the performance of MuDeRN as the second reader should be evaluated. Providing an independent second opinion could be particularly helpful when the slides were evaluated by general or less experienced pathologists.

Acknowledgments

We would like to thank contributors of BreakHis database content who kindly provided us the access to their database. We also acknowledge the University of Sydney HPC Service at the University of Sydney for providing high performance computing resources that have contributed significantly to the research results reported within this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.artmed.2018.04.005.

References

[1] Erlay J, Ervik M, Dikshit R, Eser S, Mathers C. Cancer incidence and mortality worldwide: IARC cancer base no. 11. In: GLOBOCAN 2012; 2012, v1. 0, ed.
[2] Calonge N, Petitti DB, DeWitt TG, Dietrich AJ, Gregory KD, Grossman D, et al. Screening for breast cancer: US preventive services task force recommendation statement. Ann Intern Med 2009;151:716–26.
[3] Weaver DL, Rosenberg RD, Barlow WE, Ichikawa L, Carney PA, Kerlikowske K, et al. Pathologic findings from the breast cancer surveillance consortium: population-based outcomes in women undergoing biopsy after screening mammography. Cancer 2006;106:732–42. Feb 15.
[4] Allison KH, Reisch LM, Carney PA, Weaver DL, Schnitt SJ, O'malley FP, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. Histopathology 2014;65:240–51.
[5] Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama 2015;313:1122–32.
[6] Khazai L, Middleton LP, Goktepe N, Liu BT, Sahin AA. Breast pathology second review identifies clinically significant discrepancies in over 10% of patients. J Surg Oncol 2015;111:192–7.
[7] Gandomkar Z, Brennan PC, Mello-Thoms C. Computer-based image analysis in breast pathology. J Pathol Inf 2016;7.
[8] Weyn B, van de Wouwer G, van Daele A, Scheunders P, van Dyck D, van Marck E, et al. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. Cytometry 1998;33:32–40.
[9] Filipczuk P, Kowal M, Obuchowicz A. Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology. In: Engineering in medicine and biology society (EMBC), 2013 35th Annual international conference of the IEEE; 2013. p. 7368–71.
[10] Yang L, Chen W, Meer P, Salaru G, Goodell LA, Berstis V, et al. Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. IEEE Trans Inf Technol Biomed 2009;13:636–44.
[11] Barroso-Sousa R, Metzger-Filho O. Differences between invasive lobular and invasive ductal carcinoma of the breast: results and therapeutic implications. Ther Adv Med Oncol 2016;8:261–6.
[12] Guray M, Sahin AA. Benign breast diseases: classification, diagnosis, and management. The Oncologist 2006;11:435–49.
[13] Cserni G, Orosz Z, Kulka J, Sápi Z, Kálmán E, Bori R. Divergences in diagnosing nodular breast lesions of noncarcinomatous nature. Pathol Oncol Res 2006;12:216–21.
[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977:159–74.

[15] Lawton TJ, Acs G, Argani P, Farshid G, Gilcrease M, Goldstein N, et al. Interobserver variability by pathologists in the distinction between cellular fibroadenomas and phyllodes tumors. Int J Surg Pathol 2014;22:695–8, 08/26.
[16] Longacre TA, Ennis M, Quenneville LA, Bane AL, Bleiweiss IJ, Carter BA, et al. Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. Mod Pathol 2006;19:195.
[17] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[18] Lo S-C, Lou S-L, Lin J-S, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans Med Imaging 1995;14:711–8.
[19] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–105.
[20] Zilly J, Buhmann JM, Mahapatra D. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Comput Med Imag Graph 2017;55:28–41.
[21] Lu F, Wu F, Hu P, Peng Z, Kong D. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. Int J Comput Assisted Radiol Surg 2017;12:171–82.
[22] Yu L, Yang X, Chen H, Qin J, Heng P-A. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In: AAAI; 2017. p. 66–72.
[23] Zhang Q, Xiao Y, Dai W, Suo J, Wang C, Shi J, et al. Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics 2016;72:150–7.
[24] Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys 2016;43:6654–66.
[25] Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 2015;108:214–24.
[26] Ngo TA, Lu Z, Carneiro G. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Med Image Anal 2017;35:159–71.
[27] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. Journal Pathology Informatics 2016;7.
[28] Xu Z, Huang J. Detecting 10,000 cells in one second. In: International conference on medical image computing and computer-assisted intervention; 2016. p. 676–84.
[29] Song Y, Zhang L, Chen S, Ni D, Lei B, Wang T. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans Biomed Eng 2015;62:2421–33.
[30] Sirinukunwattana K, Pluim JP, Chen H, Qi X, Heng P-A, Guo YB, et al. Gland segmentation in colon histology images: the glass challenge contest. Med Image Anal 2017;35:489–502.
[31] Drozdzal M, Chartrand G, Vorontsov E, Shakeri M, Di Jorio L, Tang A, et al. Learning normalized inputs for iterative estimation in medical image segmentation. Med Image Anal 2018;44:1–13.
[32] Sirinukunwattana K, Raza SEA, Tsang Y-W, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196–206.
[33] Kim E, Corte-Real M, Baloch Z. A deep semantic mobile application for thyroid cytopathology. In: Medical imaging 2016: PACS and imaging informatics: next generation and innovations; 2016. p. 97890A.
[34] Rezaeilouyeh H, Mollahosseini A, Mahoor MH. Microscopic medical image classification framework via deep learning and shearlet transform. J Med Imaging 2016;3. p. 044501.
[35] Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: International conference on medical image computing and computer-assisted intervention; 2013. p. 411–8.
[36] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using convolutional neural networks, in neural networks (IJCNN). In: International joint conference on, 2016; 2016. p. 2560–7.
[37] Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Medical imaging 2014: digital pathology; 2014. p. 904103.
[38] Bejnordi BE, Zuidhof G, Balkenhol M, Hermsen M, Bult P, van Ginneken B, et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. J Med Imaging (Bellingham) 2017;4. p. 044504, Oct.
[39] Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017;7.
[40] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.
[41] Todorovski L, Džeroski S. Combining classifiers with meta decision trees. Machine Learning 2003;50:223–49.
[42] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[43] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng 2016;63:1455–62.
[44] Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inf 2017;8.
[45] Kothari S, Phan JH, Moffitt RA, Stokes TH, Hassberger SE, Chaudry Q, et al. Automatic batch-invariant color segmentation of histological cancer images,biomedical imaging: from nano to macro. In: IEEE International Symposium on, 2011; 2011. p. 657–60.
[46] Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graph Appl 2001;21:34–41.
[47] Spanhol FA, Cavalin PR, Oliveira LS, Petitjean C, Heutte L. Deep features for breast cancer histopathological image classification. In: Proceedings of the IEEE conference on Systems, Man, and Cybernetics (SMC); 2017. p. 1868–73.

Please cite this article in press as: Gandomkar Z, et al. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.005