Double Transfer Learning for Breast Cancer Histopathologic Image Classification

Jonathan de Matos*,§, Alceu de S. Britto Jr.*,§, Luiz E. S. Oliveira‡ and Alessandro L. Koerich*

*École de Technologie Supérieure, Université du Québec, Montréal, QC, Canada
Email: jonathandematos@gmail.com, alessandro.koerich@etsmtl.ca
§State University of Ponta Grossa (UEPG), Ponta Grossa, PR, Brazil
†Pontificial Catholic University of Paraná, Curitiba, PR, Brazil
Email: alceu@ppgia.pucpr.br
‡Federal University of Parana, Curitiba, PR, Brazil
Email: luiz.oliveira@ufpr.com

Abstract—This work proposes a classification approach for breast cancer histopathologic images (HI) that uses transfer learning to extract features from HI using an Inception-v3 CNN pre-trained with ImageNet dataset. We also use transfer learning on training a support vector machine (SVM) classifier on a tissue labeled colorectal cancer dataset aiming to filter the patches from a breast cancer HI and remove the irrelevant ones. We show that removing irrelevant patches before training a second SVM classifier, improves the accuracy for classifying malign and benign tumors on breast cancer images. We are able to improve the classification accuracy in 3.7% using the feature extraction transfer learning and an additional 0.7% using the irrelevant patch elimination. The proposed approach outperforms the state-of-the-art in three out of the four magnification factors of the breast cancer dataset.

Index Terms—histopathologic images, transfer learning, SVM, breast cancer

I. INTRODUCTION

Cancer is characterized by an uncontrolled growth and change in a group of cells. Such an anomaly is usually associated with a part of the body (e.g. breast, rectum). The mass formed by the uncontrolled growth is called tumor and can be classified as malignant or benign. In a general definition, a benign tumor is not aggressive and does not need a rapid intervention because it does not destroy other cells and it does not spread quickly. On the other hand, a malignant tumor can be referred to as cancer and presents a destructive, invasive and fast-spreading behavior. Therefore, it is necessary fast intervention to preserve the health of the patient [1].

According to Torres et al. [2], breast cancer is the most prevalent type of cancer among women in 140 of 184 studied countries. Since 1990, the breast cancer mortality is decreasing due to the improvement in access to mammography exams and treatment. Although prevalent, in the United States 1.6 million women a year are submitted to biopsies, but only a quarter of the exams result in positive for malignant tumors. This factor increases the queue of exams, possibly delaying the start of the treatment for those that really require it. According to Gurcan et al. [6], there is a growing demand for automated methods for diagnosing. Kalinli et al. [7] presented a comparative study of computer aided diagnosis (CAD) and pathologists’ diagnosis in microscopic images of biopsies, called histopathologic images (HI). Gurcan et al. [6] raised questions related to inter-observer, intra-observer, physical, and psychological factors in the image analysis by pathologists. Inter-observer refers to the disagreement of malign and benign labeling of a sample by two pathologists. Intra-observer variations occur when a single pathologist observes the same sample within an interval of time, for example, of one day. His conclusions can be different between the two observations due to physiologic or psychological factors.

Considering the growing interest in CAD systems, Spanhol et al. [8] presented a study involving the recognition of the tumor type of breast cancer, while Kather et al. [9] presented a study involving tissue recognition for colorectal cancer. Both works present an analysis of computational techniques to automate the recognition of tumors as well as provided access to their image datasets. Several works for breast cancer classification are based on monolithic [10]–[12]. Peikari et al. [13] used a cascade of SVMs for breast cancer classification. Valkonen et al. [14] segmented whole slide images (WSI) of breast tissue using a Random Forest classifier. Balazsi et al. [15] also studied the segmentation of WSI using a Random Forest and tessellation. Araújo et al. [16] used a CNN with five convolutional layers and three fully-connected layers to classify breast cancer HI. They obtained 80.6% of overall accuracy using the CNN as a classifier and 83.3% using the
same trained network as a feature extractor and a SVM with RBF kernel as a classifier.

Since 2012 when a Convolutional Neural Network (CNN) outperformed other traditional classification methods at the ImageNet contest [17], deep approaches have also been proposed for breast cancer classification. However, the main constraints are the need of images of small size as input to limit the number of trainable parameters and the need for a large dataset to properly train such networks. Malon et al. [18] proposed feature extraction from nuclei of HI to classify mitotic cells. Bayramoglu et al. [19] proposed a custom CNN to classify breast cancer HI independently from magnification that uses low-level data augmentation to increase the size of the training dataset and image resizing. An approach using a CNN called BiCNN was proposed by Wei et al. [20] where they addressed the problem of a small size training set using data augmentation and transfer learning from the ImageNet dataset. Although producing good results, the CNN methods require data augmentation to produce more images for training as they are data-driven methods. On the other hand, the high number of samples increases the computational cost for training. Therefore, there is a trade-off between accuracy and cost.

Spanhol et al. [21] proposed an interesting approach for data augmentation using a patching procedure that has the advantage of increasing the number of images while reducing their resolution to the input size required for a pre-trained CNN. Nevertheless, such an approach uses random patches and sliding windows. The main drawback is that HI of distinct classes may have large regions with similar information, for instance, an empty space or even the stroma. Therefore, a method for improving the quality of the patches could improve the accuracy of the model.

In this paper we propose the use of a source dataset to implement a transfer learning technique for filtering patches of a target dataset. Our main hypothesis is that filtering patches based on the tissue knowledge of a colorectal cancer dataset may provide a positive impact on classifying patches of a breast cancer dataset, due to the elimination of patches that do not contain relevant information, or even that do not contribute to distinguish between types of tumor. For such an aim, in the first step, we trained a SVM on the CRC dataset to classify histopathologic images (HI) as relevant or irrelevant as illustrated in Fig. [1]. For such an aim, we use the CRC dataset which is composed of images of 150×150 pixels labeled according to the structure they contain. Eight types of structures are labeled: Tumor (T), Stroma (S), Complex Stroma (CS), Immune or lymphoid cells (L), Debris (D), Mucosa (M), Adipose (AD), and Background or Empty (E). The total number of images is 625 per structure/tissue type, resulting in 5,000 images. Fig. [2] depicts examples of the images from the dataset.

The key to the filtering process is to define what is relevant and what is not in a histopathologic image. Intuitively, only background may be considered irrelevant, but we have evaluated other possible combinations of some tissues. We have based the tissue selection for relevant and irrelevant on the labeling of the classes of the CRC dataset. The tissues are labeled from the simplest to the most important for cancer diagnosis, taking into account that even Stroma and Complex Stroma tissues contain evidence of tumor. We could have evaluated all possible combination of tissues during the relabeling procedure, but this would produce 2^8 experiments. The first step consists in mapping the patches of the CRC dataset from eight to two categories. The aim of re-labelling the images is to handle the data imbalance among categories as well as to adapt it to the filtering task. Therefore, as shown in

This paper is organized as follows. Section [II] presents the proposed approach for filtering based on the transfer knowledge from the CRC dataset to the BreaKHis dataset, as well as the transfer representation-learning from pre-trained Inception-v3 CNN. Section [III] presents the experimental results for filtering patches and classifying patients. In Section [IV] we discuss the results achieved by the proposed approach and compare them to the state-of-the-art. Finally, in the last section we present our conclusion and perspectives of future work.

II. PROPOSED METHOD

The main idea of the proposed approach is to transfer the knowledge from two source datasets to a target dataset where one of the source datasets has different data and different labels (different domain) and the other source dataset has similar data but different labels (same domain). The first dataset is ImageNet and it is used for feature representation transfer. The second dataset is CRC and it provides structural information (texture) about the tissue types of histopathologic images. Finally, the target domain is to classify histopathologic images from the BreaKHis dataset into two classes, malign and benign.

The proposed approach has two steps. In the first step, as illustrated in Fig. [1], we train a SVM to discriminate between relevant and irrelevant tissue structures in histopathologic images. In the second step, as illustrated in Fig. [5], we use this SVM to filter images from the BreaKHis dataset before training another SVM whose aim is to detect benign or malign structures into the histopathologic images.

A. Building a Filter

The idea of building a filter is to train a SVM to classify histopathologic images (HI) as relevant or irrelevant as illustrated in Fig. [1]. For such an aim, we use the CRC dataset which is composed of images of 150×150 pixels labeled according to the structure they contain. Eight types of structures are labeled: Tumor (T), Stroma (S), Complex Stroma (CS), Immune or lymphoid cells (L), Debris (D), Mucosa (M), Adipose (AD), and Background or Empty (E). The total number of images is 625 per structure/tissue type, resulting in 5,000 images. Fig. [2] depicts examples of the images from the dataset.

The key to the filtering process is to define what is relevant and what is not in a histopathologic image. Intuitively, only background may be considered irrelevant, but we have evaluated other possible combinations of some tissues. We have based the tissue selection for relevant and irrelevant on the labeling of the classes of the CRC dataset. The tissues are labeled from the simplest to the most important for cancer diagnosis, taking into account that even Stroma and Complex Stroma tissues contain evidence of tumor. We could have evaluated all possible combination of tissues during the relabeling procedure, but this would produce 2^8 experiments.

The first step consists in mapping the patches of the CRC dataset from eight to two categories. The aim of re-labelling the images is to handle the data imbalance among categories as well as to adapt it to the filtering task. Therefore, as shown in
Fig. 1. An overview of the double transfer learning from ImageNet and CRC datasets: feature extraction using two approaches (PFTAS and deep features from Inception-v3), filtering using CRC dataset as training set where TR is the training set and VL is the validation set.

Table I, we may have seven different scenarios where the total number of images mapped to our target categories, relevant (RE) and irrelevant (IR) categories, is always the same. The number of images of each source category is the same (e.g. 89 images of categories S, CS, L, D, M, AD, and E mapped to IR category for filter F1), but the ratio of the number of irrelevant images to the number of relevant images is imbalanced (e.g. 89 of S for irrelevant and 625 of T for relevant for filter F1). The imbalance is high for filters F1 and F7, but not so bad for filters F2 and F6, small for filters F3 and F5, while there is a balance for filter F4. This decision aims not to create a filter biased towards relevant or irrelevant classes.

For each scenario in Table I, we extract two different feature sets to further train a two-class SVM: PFTAS and Inception-v3 deep features. The Parameter-Free Threshold Adjacency Statistics (PFTAS) is a handcrafted texture feature extractor which counts the black pixels in the neighborhood of a pixel (between zero and eight black pixels). A nine-bin histogram registers the total count for all the pixels in the image. The images are firstly binarized by the Otsu algorithm with three thresholds at each one of the channels (Red, Green, and Blue). Additionally, the thresholded images are bitwise inverted, doubling the number of values. This results in a 162-dimensional feature vector ($9 \times 3 \times 3 \times 2$). The second feature representation is generated by an Inception-v3 CNN pre-trained with ImageNet [23]. The transfer representation learning uses the 2,048-dimensional fully-connected layer before the softmax layer as a feature vector. We used the Principal
Component Analysis (PCA) method to reduce the feature vector dimensionality from 2,048 to a lower dimension. We have chosen four target dimensions to evaluate the impact of the reduction in the accuracy: 600, 400, 200 and 100 dimensions, where 600 dimensions correspond to 95% of the accumulated variance from the most important components.

The aim of using two strategies for feature extraction is to verify if different features could improve the detection of irrelevant regions. Finally, a two-class SVM with RBF kernel is trained and evaluated for each resulting feature set for each scenario. Once the SVMs are trained, they are used for filtering the images of our target dataset, the BreakHis dataset.

B. BreakHis Patch Filtering

At the first moment, our aim is not to classify all the structures present in the images of the BreakHis dataset, but to improve the quality of the dataset before using it to train a classifier for our target task: given a HI, classify it as malign or benign. For such an aim, we use the SVMs trained at the first step, which will act as filters to eliminate irrelevant images from our target dataset.

The proposed approach relies on the fact that some images of different classes contain a high amount of information in common. Fig. 3 depicts an example of two opposite types of tumor, an Adenoma (benign) and a Ductal Carcinoma (malign) that have a similar appearance in some areas which contains the slide background. Considering that these images will be patched into square regions and that there is no prior knowledge of each tissue type inside each image, but only what type of tumor an entire image contains, regardless if the tissue structure relevant to the diagnosis occupies a large or a small region within the image.

BreakHis is the target dataset that has its images or patients recognized as malign and benign breast tumors. This dataset introduced by Spanhol et al. [8] contains 7,909 HI images of tissues stained with Hematoxylin & Eosin of 82 patients distributed in eight classes of tumors, where four classes are of malign tumors and the other four of benign tumors. The dataset is imbalanced by a factor of seven in the worst case, which means for example, that ductal carcinoma images (malign) have seven times more samples than the Adenosis (benign). Table II shows the class distribution. Each image has 700×460 pixels and for all patients there are images with four magnification factors: $40 \times, 100 \times, 200 \times$ and $400 \times$, which are equivalent to 0.49, 0.20, 0.10, 0.05 $\mu m$ per pixel respectively. Images of different magnification were obtained from the minor to the major magnification factor of a microscope when the pathologist selects a region of interest of a biopsy specimen that is processed and fixed onto glass slides. Fig. 3 shows some examples of images from the BreakHis dataset.

### TABLE II

| Tissue Type | Tumor Type    | # of Images | # of Patients |
|-------------|---------------|-------------|---------------|
| Benign      | Adenosis      | 444         | 4             |
|             | Fibroadenoma  | 1,014       | 10            |
|             | Phyllodes tumor | 453        | 3             |
|             | Tubular adenoma | 569        | 7             |
|             | Total         | 2,368       | 24            |
| Malign      | Ductal carcinoma | 3,451     | 38            |
|             | Lobular carcinoma | 626      | 5             |
|             | Mucinous carcinoma | 792     | 9             |
|             | Papillary carcinoma | 560    | 6             |
|             | Total         | 5,429       | 58            |

![Fig. 3. Example of: (a) Adenoma (benign tumor) slide; (b) Ductal Carcinoma (malignant tumor) slide](image)

First, the images of the BreakHis dataset are patched by a 150×150 rectangular sliding window with 3.3% and 8% of vertical and horizontal overlapping respectively. On average, a 700×460 image produces 15 patches. Next, we use the same feature extraction approaches presented previously, namely the PFTAS and the Inception-v3 deep features, to extract features from the patches. These feature extractors work in parallel and independently of each other. For the sake of simplicity, in Fig. 4 we represent a single block of feature extraction. Besides, the PCA is also used to reduce the dimensionality of the feature vector produced by the Inception-v3. The corresponding SVM filter trained on CRC or Inception-v3 features is used to discard irrelevant patches from the BreakHis dataset. Only the relevant patches are used then to train a two-class SVM to classify the patches of the BreakHis dataset.

The classification approach follows the protocol proposed by Spanhol et al. [8] that uses five folds with a patient-wise percentage of 70% of patients for training and 30% for testing.

![Fig. 4. Examples of images of the BreakHis dataset: Benign tumors from (a) to (d). Malign tumors from (e) to (h).](image)
Furthermore, such a protocol ensures that all images of a patient are either in the training set or in the test set, but not in both sets at the same time. Therefore, the relevant patches are split using the same protocol and the two-class SVMs are trained in the training set with 5-fold cross validation to classify patches as Benign or Malign. At the end, we aggregate the patch classification results to come up to a whole image or patient classification.

III. EXPERIMENTAL RESULTS

In the development phase, we trained the SVM filters with 85% of the CRC dataset while 15% of the images were used for validation. No test set was necessary because the purpose was not to evaluate the filter itself, but the filtering process. The selection between train and validation was random and stratified. In the development phase, we trained 35 SVM classifiers with RBF kernel using grid-search and 5-fold cross-validation for hyperparameter tuning on the data distribution shown in Table I. Seven SVMs were trained with the PFTAS features and 28 SVMs were trained with the Inception-v3 deep features, seven for each PCA dimension (100, 200, 400 and 600). The SVM filters were optimized to achieve the best accuracy for classifying relevant/irrelevant images on the validation set. Table III presents the accuracy on the validation set for SVM filters trained on the CRC dataset.

| Filter | Accuracy  |
|--------|-----------|
| PFTAS  | 95.2      |
| Inception-v3 + PCA 100 | 97.9 |
| Inception-v3 + PCA 200 | 97.9 |
| Inception-v3 + PCA 400 | 97.9 |
| Inception-v3 + PCA 600 | 98.4 |

The 7,909 images of BreaKHis dataset generated 118,635 patches of dimension 150×150 pixels. This size is the same as the images of the CRC dataset. However, keeping similar dimensions does not make the images from both datasets completely compatible, because pixels of BreaKHis’ images represent 49 µm while pixels of the CRC dataset represent 74 µm in the best case. The pixel density problem cannot be solved because it is an acquisition parameter. The patches generated from the BreaKHis dataset have a vertical overlapping of five pixels and a horizontal overlapping of 12 pixels due to the distribution of five squared patches of dimension 150×150 pixels over the 700×460 image.

The accuracy per patient is calculated by Equation (1), where \( N_{\text{correct}} \) is the number of images correctly classified of one patient and \( N_{\text{total}} \) is the total images of a patient.

\[
\text{Patient Score} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (1)
\]

The overall accuracy is obtained by Equation (2), where Patient Score comes from Equation (1) and Total Number of Patients is the number of patients.

\[
\text{Accuracy} = \frac{\sum \text{Patient Score}}{\text{Total Number of Patients}} \quad (2)
\]

All patches were submitted to the same filtering process, independently of their magnification factor. Table IV shows the number of patches, images, and patients selected as relevant after filtering and separated by the magnification factor.

A. Results for PFTAS Features

Tables V and VI present the results of the first experiments after filtering the patches of the BreaKHis dataset, only the relevant ones are used in the next step. The relevant patches are split according to the protocol proposed by Spanhol et al. [8]. Therefore, five folds were used to train the SVM classifiers with RBF kernel. The best hyperparameters \( C \) and \( \gamma \) were found using cross validation and grid-search. We executed seven experiments, one for each filter. All filters were executed separately for each magnification factor. All 28 executions of filters with magnification factors were repeated five times, resulting in 140 executions. We executed one more batch of experiments without the filter, adding five more folds for each magnification factor adding more 20 experiments. These variations summed up to 160 experiments.

TABLE IV

| Magnification | Filter | Patches | Images | Patients |
|---------------|--------|---------|--------|----------|
| 40×           | F1     | 48.7    | 10.6   | 1.7      |
|               | F2     | 100     | 73.0   | 31.5     |
|               | F3     | 95.1    | 62.7   | 24.1     |
|               | F4     | 98.7    | 75.4   | 33.5     |
|               | F5     | 100     | 99.5   | 81.3     |
|               | F6     | 93.9    | 88.4   | 70.1     |
|               | F7     | 100     | 96.9   | 90.9     |
| 200×          | F1     | 39.0    | 6.1    | 0.8      |
|               | F2     | 96.3    | 60.3   | 21.3     |
|               | F3     | 93.9    | 40.7   | 10.4     |
|               | F4     | 97.5    | 51.7   | 14.7     |
|               | F5     | 100     | 95.8   | 68.7     |
|               | F6     | 91.4    | 88.5   | 70.2     |
|               | F7     | 100     | 97.3   | 93.7     |
| 400×          | F1     | 47.5    | 9.4    | 1.3      |
|               | F2     | 100     | 61.5   | 21.3     |
|               | F3     | 92.6    | 32.8   | 7.6      |
|               | F4     | 97.5    | 41.9   | 10.1     |
|               | F5     | 100     | 87.4   | 53.3     |
|               | F6     | 97.5    | 89.0   | 59.7     |
|               | F7     | 100     | 99.8   | 97.3     |
| 850×          | F1     | 76.8    | 22.5   | 4.6      |
|               | F2     | 98.7    | 66.7   | 19.7     |
|               | F3     | 90.2    | 34.5   | 6.4      |
|               | F4     | 95.1    | 33.7   | 5.8      |
|               | F5     | 100     | 82.3   | 43.3     |
|               | F6     | 100     | 89.8   | 47.7     |
|               | F7     | 100     | 100    | 99.6     |

After filtering the patches of the BreaKHis dataset, the relevant ones are used in the next step. The relevant patches are split according to the protocol proposed by Spanhol et al. [8]. Therefore, five folds were used to train the SVM classifiers with RBF kernel. The best hyperparameters \( C \) and \( \gamma \) were found using cross validation and grid-search. We executed seven experiments, one for each filter. All filters were executed separately for each magnification factor. All 28 executions of filters with magnification factors were repeated five times, resulting in 140 executions. We executed one more batch of experiments without the filter, adding five more folds for each magnification factor adding more 20 experiments. These variations summed up to 160 experiments.

A. Results for PFTAS Features

Tables V and VI present the results of the first experiments at patch-level, image-level, and patient-level. The image-level
results are obtained by aggregating patches using majority vote and the sum rule \(^{24}\) on the probabilities predicted by the SVM classifier. The image classification results are used to calculate the patient-level accuracy.

**TABLE V**

| Magnif. | Patches | Images | Patients |
|---------|---------|--------|----------|
|        | Sum | Vote | Sum | Vote |
| 40×     | 82.9±3.4 | 85.0±4.4 | 85.4±4.4 | 86.1±4.5 | 86.4±4.7 |
| 100×    | 83.0±3.8 | 84.9±3.9 | 84.6±3.9 | 86.6±4.9 | 86.3±4.9 |
| 200×    | 86.5±3.1 | 88.3±3.5 | 88.5±3.6 | 88.4±4.6 | 88.7±4.5 |
| 400×    | 84.8±4.5 | 87.0±4.8 | 87.3±4.8 | 88.0±5.7 | 88.2±5.4 |

The results in Table [VI] consider only the patches filtered by filter F7, that considers as irrelevant only the Empty (or Background) images of CRC dataset because it produced the best results in terms of correct filtering (see Table [VII]) and it never excluded patients (see Table [V]). Only the magnification factor 200× provided improvements with filtering as highlighted in Table [VI]. However, the results of both Table [V] and [VI] surpasses the results presented in \([8]\). Spanhol et al. \([8]\) used an SVM classifier and the PFTAS feature to achieve patient-wise mean percent accuracy of 81.6±3.0, 79.9±5.4, 85.1±3.1 and 82.3±3.8 for 40×, 100×, 200× and 400× magnification respectively.

**B. Results for Inception-v3 Deep Features**

Fig. 6 presents the results of the filtering process considering the percentage of remaining patches when the Inception-v3 deep features were used instead of PFTAS features. This graph represents only the remaining patches for the magnification factor 100×. All other magnifications (40×, 200×, and 400×) follow a very similar distribution. Aggressive filters such as filters F1 to F3, consider all images from a patient as irrelevant and in this case, we have discarded these filters. Besides that, some filters may exclude all patches from an image, but this is not a problem, meaning that the excluded image does not contribute to the classification.

**TABLE VI**

| Magnif. | Patches | Images | Patients |
|---------|---------|--------|----------|
|        | Sum | Vote | Sum | Vote |
| 40×     | 82.5±3.3 | 85.1±4.6 | 84.9±5.2 | 86.4±4.3 | 86.1±4.9 |
| 100×    | 82.7±3.6 | 85.4±3.9 | 84.8±3.6 | 87.0±4.3 | 86.6±4.1 |
| 200×    | 86.7±3.6 | 89.0±3.9 | 88.9±4.3 | 89.2±5.1 | 89.3±5.3 |
| 400×    | 84.8±4.6 | 87.0±4.9 | 87.2±5.0 | 87.9±5.7 | 88.2±5.8 |

Fig. 6. Percentage of remaining patches (relevant) for each filter using Inception-v3 deep features.

Table VII shows only the results at patient-level by aggregating patches using the sum rule and considering the two best filters (F6 or F7) or no filtering. In general, the mean percent accuracy for three out of four magnifications is better than those achieved with the PFTAS features. It is also possible to see that the filtering process only loses in the magnification factor of 40× for the feature vector of dimension 200.

Fig. 7 shows the relation between wins and losses considering the filtered and non-filtered executions for each fold. We can notice that the filtering process produces worst results only in fold 1 for all 16 executions presented in Table VII. For fold 1, it loses 13 times, but for the other folds, there are much more wins than losses. This result is related to the small number of patients for some categories of tumor, what makes some patients present in the test set of fold 1 and not present in the test set of other folds. There are only three patients with Phyllodes tumor. Therefore, when the fold is split into training (70%) and test (30%) only one patient remains in the test set. A similar situation occurs to the Adenosis and Lobular Carcinoma categories, which have only four and five patients respectively. Even if these two categories of tumor...
have a slightly higher number of patients, the imbalance also affects the performance. The images filtered in these test sets may not react well to the knowledge of the CRC dataset. Our aim is not only augmenting the amount of data but also improving the patch quality by reducing the number of similar regions that may be present in different categories of tumor. Table VIII shows that, except for the 40× magnification, the proposed approach is the most accurate.

It is worth mentioning that the results presented by Han et al. [25] outperform all the results reported in Table VIII achieving 97.1%, 95.7%, 96.5% and 95.7% for the four magnification factors. However, such results are not comparable because they have used a completely different experimental protocol that did not respect the folder distribution proposed by Spanhol et al. [8].

We proposed a patch procedure with filtering to improve the quality of images for training a SVM classifier for breast cancer HI classification. The proposed approach uses two steps of transfer learning, one at feature extraction level, where deep features are extracted using an Inception-v3 CNN pre-trained with ImageNet dataset and the other at tissue structure level, where tissue images from another HI dataset are used to filter out irrelevant tissue structures of our target dataset. The experimental results have shown that the proposed filtering approach using transfer learning outperforms previous approaches.

### Table VII

| PCA Dimension | Magnification | Filter  | Filter  | No Filter |
|---------------|---------------|---------|---------|-----------|
| 100           | 40×           | F6      | 89.9±3.6| 88.5±3.8  |
| 100×          | F7            | 91.0±3.0| 90.3±3.4|           |
| 200           | 40×           | F6      | 89.7±3.6| 88.6±3.6  |
| 200×          | F7            | 86.7±1.6| 85.7±2.4|           |
| 400           | 40×           | F6      | 88.7±2.9| 89.0±3.4  |
| 400×          | F7            | 90.6±3.4| 90.0±3.7|           |

The main drawback of the proposed approach is the computational cost for training the SVM classifier due to the number of patches that we produced as well as due to dimensionality of the feature vectors. The number of instances impacts directly in the size of the kernel matrix and the dimensionality of the features impacts in the computation cost of the matrix calculation. The feature extraction is relatively time-consuming due to the size of the Inception-v3 CNN, but GPU computation helps to speed this step. Although our approach is costly, the approaches that used only CNNs are much more expensive, taking more than 40 minutes for a single experiment, as reported in [21].

Another drawback of the proposed approach is the number of losses in fold 1, as shown in Fig. 7. However, further analysis is required to detect what causes such losses in this particular fold. Therefore, we believe that there is still some room for an overall improvement if we could improve the performance for this fold.

### Table VIII

| Method                  | Magnification |
|-------------------------|---------------|
|                         | 40×           | 100×          | 200×          | 400×          |
| Proposed (PFTAS + Filter)| 86.4 86.3 88.7| 88.2           |               |
| Proposed (Inception-v3) | 86.1 89.3     | 85.7           |               |
| Proposed (PFTAS)        | 88.5 90.3     | 86.9           |               |
| Proposed (Inception-v3) | 88.9 89.7     | 86.7           |               |
| Baseline [8]            | 81.6 79.6     | 85.1 82.7      |               |
| CNN (Alexnet) [21]      | 90.0 88.4     | 84.6 86.1      |               |
| Deep Features (DeCaf) [26] | 83.6 83.8 86.3 | 82.1          |
| CNN + Fisher [27]       | 90.0 88.9     | 86.9 86.3      |               |
| MI Approach [28]        | 92.1 89.1     | 87.2 82.7      |               |

The proposed filtering and patching approaches improved the classification results in comparison of using the whole image [8]. Table VIII presents a list of state-of-the-art methods that were evaluated on the BreaKHis dataset using similar experimental protocols. Spanhol et al. [21] proposed a random patching procedure with the purpose of providing data augmentation for training a CNN, where random patches of sizes 64×64 and 32×32 pixels are generated by an overlapping sliding window. The number of patches generated by such a strategy is much higher than the proposed approach. However,
The proposed approach also helped in solving the problem of scarcity of data besides improving the quality of the patch images due to the SVM classifier that learns from features extracted by other methods.

The results achieved by the proposed approach have shown that it is possible to exploit the knowledge from other datasets to improve the accuracy of a particular classification task. A common problem in HI is the lack of samples due to the expensive process of acquisition, analysis and labeling of images. Therefore, we used the knowledge acquired from the ImageNet dataset to extract meaningful features. Despite the differences between the images of the ImageNet and HI datasets, the Inception-v3 CNN performed well.

Another possible approach could be training from scratch using other HI datasets. However due to the low number of samples in such a kind of datasets, training a CNN is not feasible. To circumvent this problem, the fine-tuning is a way of reusing the representation learned from other tasks or datasets and adapt them to a new context by retraining only a small part of a CNN. We intend to use this approach in future works. The knowledge transferred from the CRC dataset was useful and it provided some improvement. In the future, we intend to use the activation layers of a CNN trained on other HI datasets to extract patches based on class-activated regions. Our idea is that activation regions with higher values have a strong contribution to the final classification.

REFERENCES

[1] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Systems with Applications, vol. 36, no. 1, pp. 3240–3247, 2009.

[2] L. A. Torre, F. Islam, R. L. Siegel, E. M. Ward, and A. Jemal, “Global cancer in women: Burden and trends,” CEBP Focus: Global Cancer in Women, vol. 26, no. 4, pp. 444–457, 2017.

[3] G. N. Hortobagyi, J. de la Garza Salazar, K. Pritchard, D. Amadori, R. Huidinger, C. A. Hudis, H. Khaled, M.-C. Li, M. Martin, M. Namer, J. A. O'Shaughnessy, Z. Z. Shen, and K. S. Albin, “The Global Breast Cancer Burden: Variations in Epidemiology and Survival,” Clinical Breast Cancer, vol. 6, no. 5, pp. 391–401, 2005.

[4] J. Soler, J. Navarro, N. Garcia, and F. Amoros, “Ultrasound based application for intraglandular mapping of breast cancer,” Computer Methods and Programs in Biomedicine, vol. 112, no. 2, pp. 293–301, 2013.

[5] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. O'Malley, and D. L. Weaver, “Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens,” J. of American Medical Association, vol. 313, no. 11, pp. 1122–1132, 2015.

[6] M. N. Gurcan, L. E. Boucheron, and A. Can, “Histopathological Image Analysis: A Review,” IEEE Rev. Biomed. Eng., vol. 2, no. 1, pp. 147–171, 2009.

[7] A. Kalinli, F. Sarikoc, H. Akgun, and F. Ozturk, “Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples,” Computers Methods and Programs in Biomedicine, vol. 110, no. 3, pp. 298–307, 2013.

[8] F. A. Spathol, L. S. Oliveira, C. Petitjean, and L. Heute, “A Dataset for Breast Cancer Histopathological Image Classification,” IEEE Trans. Biomed. Eng., vol. 63, no. 7, pp. 1455–1462, 2016.

[9] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Guiser, A. Marx, and F. G. Zöllner, “Multi-class texture analysis in colorectal cancer histology,” Scientific Reports, vol. 6, no. 1, p. 27988, 2016.

[10] S. A. Korkmaz and M. Poyraz, “Least Square Support Vector Machine and Minimum Redundancy Maximum Relevance for Diagnosis of Breast Cancer from Breast Microscopic Images,” Procedia - Social and Behavioral Sciences, vol. 174, pp. 4026–4031, 2015.

[11] H. Irshad, A. Gouaillard, L. Roux, and D. Racoeceanu, “Multispectral band selection and spatial characterization: Application to mitosis detection in breast cancer histopathology,” Computerized Medical Imaging and Graphics, vol. 38, no. 5, pp. 390–402, 2014.

[12] A. Tashk, M. S. Helfroush, H. Danyali, and M. Akbarzadeh-jahromi, “Automatic detection of breast cancer mitotic cells based on the combination of textural, statistical and innovative mathematical features,” Applied Mathematical Modelling, vol. 39, no. 20, pp. 6165–6182, 2015.

[13] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, “Automatic cellularity assessment from post-treated breast surgical specimens,” Cytometry Part A, vol. 91, no. 11, pp. 1078–1087, 2017.

[14] M. Valkonen, K. Karatsalo, K. Liimatainen, M. Nykyter, L. Latonen, and P. Ruusuvuori, “Metastasis detection from whole slide images using local features and random forests,” Cytometry Part A, vol. 91A, no. 6, SI, pp. 555–565, jun 2017.

[15] M. Balazsi, P. Blanco, P. Zoroquian, M. D. Levine, and M. N. Burmeier Jr., “Invasive ductal breast carcinoma detector that is robust to image magnification in whole digital slides,” J. of Medical Imaging, vol. 3, no. 2, 2016.

[16] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguilar, C. Eloy, A. Polonia, and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” PLOS ONE, vol. 12, no. 6, pp. 1–14, 06 2017.

[17] A. Kritchevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.

[18] C. Malon, E. Brachtel, E. Cosatto, H. P. Graf, A. Kurata, M. Kuroda, J. S. Meyer, A. Saito, S. Wu, and Y. Yagi, “Mitotic figure recognition: Agreement among pathologists and computerized detector,” Analytical Cellular Pathology, vol. 35, no. 2, pp. 97–100, 2012.

[19] N. Bayramoglu, J. Kannala, and J. Heikilä, “Deep learning for magnification independent breast cancer histopathology image classification,” in 23rd Intl Conf. on Pattern Recognition, Dec 2016, pp. 2440–2445.

[20] B. Wei, Z. Han, X. He, and Y. Yin, “Deep learning model based breast cancer histopathological image classification,” in IEEE 2nd Intl Conf. on Cloud Computing and Big Data Analysis, April 2017, pp. 348–353.

[21] F. A. Spathol, L. S. Oliveira, C. Petitjean, and L. Heute, “Breast cancer histopathological image classification using convolutional neural networks,” in 2016 Intl Joint Conf. on Neural Networks, July 2016, pp. 2560–2567.

[22] L. P. Coelho, “Mahotas: Open source software for scriptable computer vision,” Cytometry Part A, vol. 7, no. 4, pp. 249–252, 2013.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” CoRR, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567.

[24] J. Kittert, M. Hatef, R. P. W. Dun, and J. Matas, “On combining classifiers,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pp. 229–239, March 1998.

[25] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, “Breast cancer multi-classification from histopathological images with structured deep learning model,” Scientific Reports, vol. 7, no. 4172, pp. 1–10, June 2017.

[26] F. A. Spathol, L. S. Oliveira, P. R. Cavanil, C. Petitjean, and L. Heute, “Deep features for breast cancer histopathological image classification,” in IEEE Intl Conf. on Systems, Man, and Cybernetics, Oct 2017, pp. 1868–1873.

[27] Y. Song, J. J. Zou, H. Chang, and W. Cai, “Adapting fisher vectors for histopathological image classification,” in IEEE 14th Intl Symposium on Biomedical Imaging, April 2017, pp. 600–603.

[28] P. J. Sadrsharshan, C. Petitjean, F. Spathol, L. S. Oliveira, L. Heute, and P. Honeine, “Multiple instance learning for histopathological breast cancer image classification,” Expert Systems with Applications, vol. 117, pp. 103 – 111, 2019.