

# Patch-based system for Classification of Breast Histology images using deep learning

Kaushiki Roy\*, Debapriya Banik, Debotosh Bhattacharjee, Mita Nasipuri

*Department of Computer Science and Engineering, Jadavpur University, Kolkata-32, India*



## ARTICLE INFO

### Article history:

Received 5 April 2018

Received in revised form

26 November 2018

Accepted 28 November 2018

### Keywords:

Histopathological breast images

patch-based classifier

deep learning

convolutional neural networks

majority voting

## ABSTRACT

In this work, we proposed a patch-based classifier (PBC) using Convolutional neural network (CNN) for automatic classification of histopathological breast images. Presence of limited images necessitated extraction of patches and augmentation to boost the number of training samples. Thus patches of suitable sizes carrying crucial diagnostic information were extracted from the original images. The proposed classification system works in two different modes: one patch in one decision (OPOD) and all patches in one decision (APOD). The proposed PBC first predicts the class label of each patch by OPOD mode. If that class label is the same for all the extracted patches and that is the class label of that image, then the output is considered as correct classification. In another mode that is APOD, the class label of each extracted patch is extracted as done in OPOD and a majority voting scheme takes the final decision about class label of the image. We have used ICIAR 2018 breast histology image dataset for this work which comprises of 4 different classes namely normal, benign, in situ and invasive carcinoma. Experimental results show that our proposed OPOD mode achieved a patch-wise classification accuracy of 77.4% for 4 and 84.7% for 2 histopathological classes respectively on the test set obtained by splitting the training dataset. Also, our proposed APOD technique achieved image-wise classification accuracy of 90% for 4-class and 92.5% for 2-class classification respectively on the split test set. Further, we have achieved accuracy of 87% on the hidden test dataset of ICIAR-2018.

© 2018 Elsevier Ltd. All rights reserved.

## 1. Introduction

Breast cancer is one of the leading cause ([Rangayyan et al., 2007](#)) of cancer in women worldwide. According to a report published by WHO ([World Health Organization, 2018](#)) in 2013, nearly five lakh women lost their lives due to this deadly disease worldwide in 2011. In India, too, the number of breast cancer incidences are rising at an alarming rate. It has now become the most common cancer among women in most cities in India and the second most common disease in rural areas. In addition, most females diagnosed with breast cancer are in the younger age group (25–40 years). The risk of breast cancer ([Surakasula et al., 2014](#)) increases alarmingly until menopause then it decreases gradually. Breast cancer diagnosis consists of a series of steps. Whenever a lump or nodule is discovered in a breast during clinical examination, various screening tests like mammography ([Behrens et al., 2007](#)) or ultrasound is performed to detect changes in the breast. These screening tests are

followed by a biopsy to make a definite diagnosis and detect any malignant growth in the breast tissue. Biopsies enable a doctor to analyze the microscopic structure of the tissue and hence differentiate between normal, benign or malignant lesions and accordingly, performs successive prognosis. Cancer might become fatal if not detected early. However, early detection of this deadly disease can decrease the mortality rate since more treatment options become available when discovered in the early stage. The traditional method of inspecting the biopsy slides under the microscope is laid on the shoulders of the pathologists. However, this manual inspection technique is time-consuming and is dependent on the expertise of the pathologists. Thus developing an automated system for breast cancer detection from the breast histology images is the need of the hour. Carcinomas can be divided into two classes namely in-situ and invasive carcinomas. An in-situ carcinoma is one in which the malignant growth is restricted to the tissues in which they have occurred and have not spread to the surrounding tissues. In contrary, an invasive carcinoma is one in which the malignant growth has spread to surrounding areas from their point of occurrence. The primary task in developing an automated system for breast cancer detection is to classify the breast histology images into four class namely normal, benign, in situ or invasive carcinomas. In this paper, we have reviewed some of the recent

\* Corresponding author at: Department of Computer Science and Engineering, Jadavpur University, Kolkata-32, India.

E-mail addresses: [kaushiki.cse@gmail.com](mailto:kaushiki.cse@gmail.com) (K. Roy), [debu.cse88@gmail.com](mailto:debu.cse88@gmail.com) (D. Banik), [debotoshb@hotmail.com](mailto:debotoshb@hotmail.com) (D. Bhattacharjee), [mitanasipuri@gmail.com](mailto:mitanasipuri@gmail.com) (M. Nasipuri).

state-of-the-art techniques for automated breast histology classification and have developed a patch-based classifier (PBC) using deep learning approach for automated classification of breast histology images.

### 1.1. Motivation

As mentioned in the previous section, breast cancer is one of the deadliest diseases amongst women worldwide, and the traditional method of microscopic inspection is highly time-consuming and prone to manual errors. This motivated us to develop an automated system for classification of microscopic breast histopathological images. Most of the reported literature for this work use hand-crafted features for breast cancer classification. However, deep learning approaches eliminate the need for extracting hand-crafted features. Thus, in this work, we have developed a patch-based classifier (PBC) using the convolutional neural network (CNN) for automated breast histology image classification. The details of the work are mentioned in section 2.

### 1.2. Related Works

This section describes some of the standard state-of-the-art methodologies for breast cancer detection from histopathological images. The state-of-the-art can be broadly categorized as either hand-crafted feature-based approach or deep learning based approach using the convolutional neural network (CNN).

The hand-crafted features (He et al., 2012) used by most researchers are mostly thresholding-based, clustering-based, active contour-based, watershed-based, graph-cut, etc. The hand-crafted features mainly aim at segmenting the nuclei from the entire breast cancer (BCa) histopathology slide images. Distinguishing features are extracted from the segmented nuclei to differentiate between malignant and benign slides. In (Veta et al., 2013), the fast radial symmetry-based approach followed by marker-controlled watershed segmentation was used for nuclei extraction from breast cancer histopathology images. In their work, 39 biopsy slide images were acquired from 38 different patients. In another work (Jain et al., 2014), Chan Vese (CV) model based active contour technique was implemented to segment the cells from the background in breast histopathological images. Morphological features were extracted from these segmented cells which were used for classification of the cells as either normal or cancerous. In (Basavanhally et al., 2013), geodesic based active contour model was used for segmenting nuclei from BCa histopathological images. Both architectural and textural features of the nuclei were considered. Graph-based features (architectural feature) and 13 Haralick features (texture features) were extracted from the segmented nuclei which were used for classification and developing an automated system for detecting the Modified Bloom–Richardson (mBR) (*Breast cancer and breast pathology*, 2018) grade of different histopathological slides. In their work, breast histopathological images collected from 126 different patients were considered. In (Khan et al., 2013), an automated system for segmenting tumor cells in the BCa histopathological images by segmenting the image into hypocellular and hypercellular stroma regions using magnitude and phase spectra in the frequency domain was proposed. They have worked on MITOS dataset (MITOS Dataset, 2018) which consist of 35 breast histopathological images collected from 5 different patients. In (Roullier et al., 2016), graph-based segmentation has been used to extract the mitotic nuclei from the BCa histopathological whole slide images (WSI). In (Kaymak et al., 2017), an artificial neural network (ANN) based approach has been used for automatic breast histology classification.

In recent years, Convolutional Neural Networks (CNNs) has gained immense importance for breast histopathological image

classification. CNNs have huge advantages over the hand-crafted feature extraction techniques since CNNs extract features automatically from the image patches and the results obtained are comparable with those obtained from traditional feature extraction techniques. Spanhol et al. (Fabio Alexandre Spanhol et al., 2016) have worked on BreakHis dataset which consists of microscopic histopathological breast images captured at different magnifications. They have developed a CNN model to classify the images as either benign or malignant. The authors reported that the accuracy of the system decreased with increase in magnification since at higher magnification their CNN architecture failed to extract useful features. In another work (Cruz-Roa et al., 2014), a CNN model was proposed for the automatic classification of invasive ductal carcinoma in whole slide images (WSI) and hence to differentiate between the invasive and non-invasive images. Both (Fabio Alexandre Spanhol et al., 2016) and (Cruz-Roa et al., 2014) is a 2-class classification problem where the classes are either benign/malignant or invasive/ non-invasive. In (Wahab et al., 2017), authors have proposed a CNN model for separating the mitotic and non-mitotic nuclei from breast histopathological images. In (Vang et al., 2018), the authors have used the pre-trained Inception-V3 model (Szegedy et al., 2016) for 4 class classification of breast cancer histopathology image with some post-processing techniques. The Inception-V3 model is a pre-trained model that was developed for the classification of the images in the ImageNet database into 1000 different image classes.

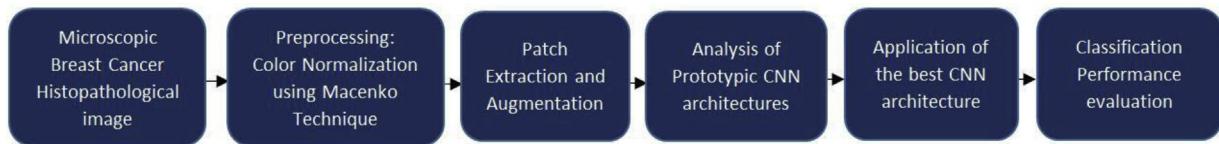
However, the field of deep learning has been very less explored in the field of breast cancer histopathological image classification. The few state-of-the-art that exists performs 2 class classification that is a classification of the histopathological images into two histological classes namely normal and malignant. However, none of the reported state-of-the-art separates the benign ones from the normal ones. In addition, malignancy is also of two types *in-situ* (cancer cells are limited to the regions in which they have occurred) and *invasive* (cancer cells have spread to the surrounding tissues from their point of occurrence). Thus to develop a fully automated system for histopathological image classification, all the different categories should be considered. Further, deep learning models eliminate the need for extracting hand-crafted features for performing automatic classification and outperforms the results obtained with hand-crafted features in most cases. The performance of automated systems for classification using hand-crafted features is mainly dependent on nuclei segmentation step as in (Veta et al., 2013), (Jain et al., 2014), (Basavanhally et al., 2013). But the performance of deep learning approaches is not limited by the classification results of nuclei segmentation step since training and classification using deep learning is based on the direct processing of image regions. This motivated the authors in this paper to develop a fully automatic system for 4-class (normal, benign, *in situ* and invasive carcinoma) and 2-class breast cancer histopathological image classification using deep learning approaches.

In this work, we have proposed a patch-based classifier (PBC) using CNN to classify breast histopathological images into four as well as two histopathological classes. The details of the classes and the CNN architecture deployed for this work is mentioned in section 2.5.

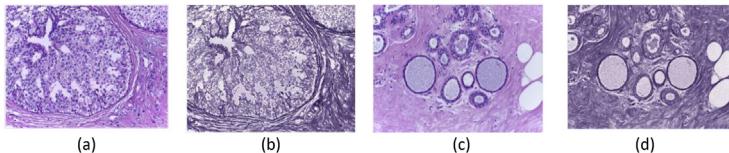
#### 1.2.1. Contribution

The main contribution of this paper can be summarized as follows-

- In this work, we have developed a patch-based classifier (PBC) which uses an optimal architecture of a convolutional neural network (CNN), for automated classification of breast cancer histopathology images.



**Fig. 1.** Block Diagram of the proposed methodology.



**Fig. 2.** (a) and (c) represents H&E stained histopathological biopsy slides. (b) and (d) represent images obtained after stain normalization corresponding to (a) and (c) respectively.

- The proposed classification system works in two different modes: one patch in one decision (OPOD) and all patches in one decision (AOPD). The patch labels are predicted by OPOD mode, and the result is obtained unanimously whereas in the AOPD mode class label of the image is obtained by a majority voting scheme.
- To verify the classification ability of the proposed system, the breast histopathological images are classified into 2 classes (non-malignant and malignant) as well as 4 classes (normal, benign, in situ and invasive carcinoma) while most of the existing methods classify the same broadly into 2 classes.
- We have also explored the potentiality of our proposed model in classifying the images in the test dataset obtained by splitting the training set as well as the actual hidden test data set of ICIAR-2018 breast cancer histology image dataset.
- Our model achieves an accuracy of 87% in classifying the images of ICIAR-2018 hidden test dataset.

This paper is organized as follows. Section 2 describes elaborately the materials and methodology employed in this paper. Section 3 contains the experimental results, discussions, and comparison with the state-of-the-art. Finally, the paper is concluded by section 4.

## 2. Materials and Methods

### 2.1. Schematic representation

Fig. 1 represents the block diagram of the entire methodology used in this work. Each of the parts is explained elaborately in the upcoming sections.

### 2.2. Preprocessing

For examination of histopathological slides, stains are used to enhance the contrast between the different histological structures especially the nuclei and the cytoplasm which eases their manual inspection under the microscope. The most commonly used stain in histopathological slides for their microscopic examination is the hematoxylin and eosin (H&E) stain. Hematoxylin colors the nuclei with a bluish shade whereas eosin gives a reddish pink color to the cytoplasm. Preprocessing (Azevedo Tosta et al., 2017) of the H&E stained images is necessary before developing an automated system for histopathological image analysis. H&E stain images are prone to undesirable color variation since different slide scanners respond differently to various colors. Also, slides prepared differently by using different stain concentrations absorb different amount of lights which results in different appearance. However, standardization of image appearance is necessary to develop an

automated system for histopathological image analysis. Color normalization is a technique which can reduce such variations. In our work, Macenko (Macenko et al., 2009) color normalization technique has been used. The colors in the input images were transformed to optical density with logarithmic function (Vang et al., 2018) to normalize their concentration given in equation 1.

$$OD = -\log \left( \frac{I}{I_0} \right) \quad (1)$$

Here  $I$  is the intensity of the R, G, B channel if the RGB color model is used and  $I_0$  is the intensity of the channel in the absence of any stain.

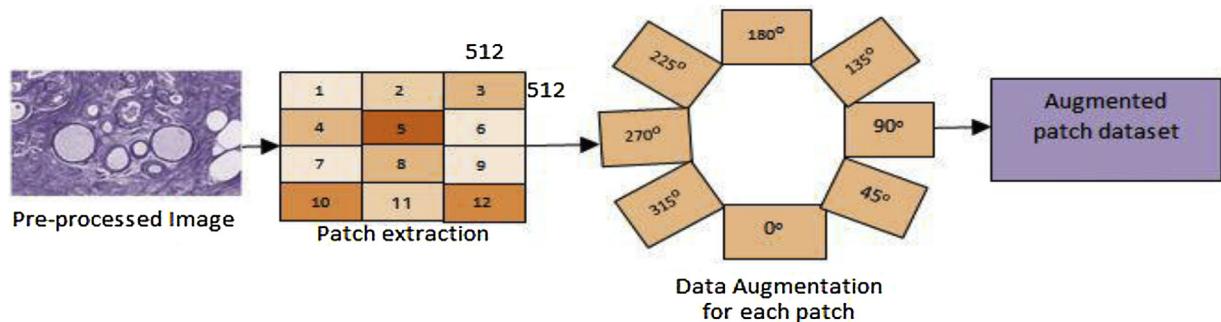
Generally, stained images are of higher dimension, and for the implementation of effective image processing techniques and restoration of useful features, singular value decomposition (SVD) (Khan et al., 2014) is commonly applied on the OD vectors which project the image in a 2D plane preserving relevant features by considering vectors with a higher order of variance. The original images and the ones obtained after normalization is shown in Fig. 2.

### 2.3. Patch extraction and Augmentation

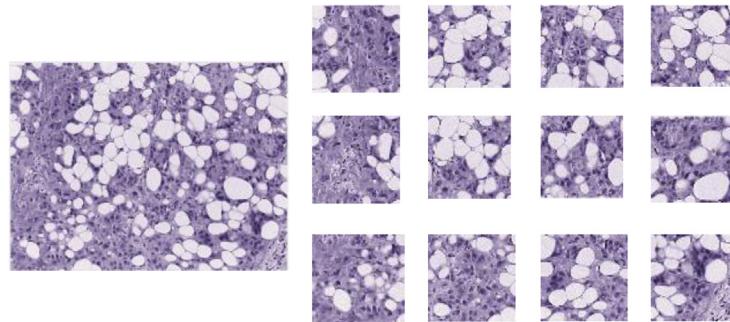
The performance of a CNN network depends on the number of training examples available and as a rule of thumb more is the number of training samples (Deng and Dong, 2014) available per class, more discriminating feature amongst the various classes can be extracted and hence the classification accuracy increases. However, our dataset had a limited number of training samples available for each category, so patch extraction and augmentation was necessary to boost the number of training samples per class. The block diagram for patch extraction and augmentation from a preprocessed image is shown in Fig. 3. The extracted and the augmented patches are finally stored in the augmented patch dataset.

Thus our developed histopathological slide classification framework consists of an automatic non-overlapping patch extraction step. Let  $I^{H \times W}$  be the input preprocessed image and  $H$ ,  $W$  be its height, and width respectively. This image is divided into  $n$  number of  $P_i$  patches,  $i = \{1, 2, \dots, n\}$  each having dimensions  $h \times w$  and  $n = (H/h) \times (W/w)$ .

Our input images have size  $2048 \times 1536$ . Individual non-overlapping square sized patches of size  $512 \times 512$  were extracted from each image resulting in a total of 12 patches from a single image as shown in Fig. 4. Nuclei structure (Irshad et al., 2014) plays a very crucial role in analyzing a histopathological image. Since a nucleus has a roughly circular shape having equal radii in all direction, so square-sized patches were extracted from each image. Now an obvious question arises why the non-overlapping patches were extracted? The reason being overlapping patches contains redundant information which can lead to overfitting in the CNN model.



**Fig. 3.** Block Diagram of patch extraction and augmentation from a sample preprocessed image.



**Fig. 4.** Input preprocessed image and the twelve patches extracted from it.

**Table 1**  
List of Abbreviations.

Abbr.	Description	Abbr.	Description
CNN	Convolutional Neural Network	rmsprop	Root mean square propagation
PBC	Patch-Based Classifier	SVD	Singular Value Decomposition
APOD	All Patches in One Decision	MNR	Macenko Normalized
OPOD	One Patch in One Decision	WSI	Whole Slide Images
CV	Chan Vese	mBR	Modified Bloom Richardson

dant information for one or more pixels, so to avoid redundancy we have used non-overlapping patches.

It is worth mentioning that we have avoided the use of smaller patches of dimension  $64 \times 64$  or  $128 \times 128$ . This is because in ICIAR 2018 challenge dataset (ICIAr, 2018), the label has been assigned to the whole input histopathological images of size  $2048 \times 1536$  and there is no guarantee that smaller squared patches of size  $64 \times 64$  or  $128 \times 128$  would carry sufficient diagnostic information. Also, a larger patch of size  $512 \times 512$  would provide a larger field of view (FOV) and carry more discriminating features in contrast to the smaller patches as illustrated in Fig. 5.

From Fig. 5, it is evident that for two different classes of images as shown in (5a) and (5e), smaller patches of size  $64 \times 64$  and  $128 \times 128$  in (5b), (5c), (5f) and (5g) have smaller field of view (FOV) and does not contain any discriminating features. However, in patches of size  $512 \times 512$  the FOV is much larger and contain sufficient discriminating features as shown in (5d) and (5 h). Moreover, the detailed diagnostic information about the nuclei, their surrounding tissue structure as in (5d) and (5 h) is required to classify images into different classes (Table 1).

In addition, we have performed an experimental analysis of learning accuracy obtained when the proposed PBC was trained with different patch sizes ( $64 \times 64$ ,  $128 \times 128$  and  $512 \times 512$ ) and

**Table 2**  
Comparative analysis of different patch sizes and learning accuracy.

Patch Size	Learning Accuracy (%)
$64 \times 64$	51
$128 \times 128$	58.5
$512 \times 512$	85.8

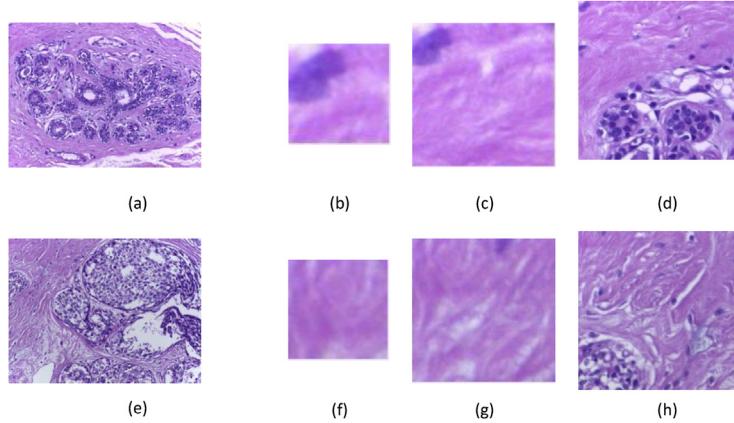
the same is reported in Table 2. We have observed that the learning accuracy with patch size  $512 \times 512$  was much higher than the accuracy obtained with smaller patches.

Now one may be inquisitive to know why a larger patch of size, say,  $1024 \times 1024$  was avoided? The maximum size of non-overlapping square sized patches that can be obtained from our input image of size  $2048 \times 1536$  is  $512 \times 512$ . Moreover, it is noted that the square sized patches of  $512 \times 512$  were sufficient to cover the essential tissue structures in a histopathological image.

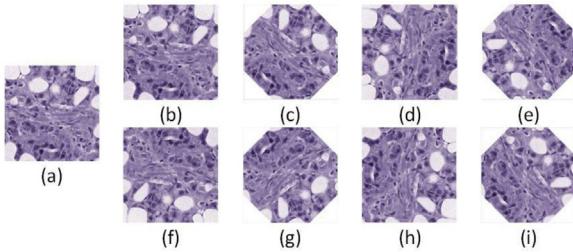
Dataset augmentation (Krizhevsky et al., 2012) is an integral part of deep learning since it helps to overcome overfitting on models by increasing the number of training samples especially in cases where the number of samples in the original dataset is low. Dataset augmentation (Krizhevsky et al., 2012) is utilized to expand the training set by domain-specific transformations through translation, rotation, and flipping. Medical images are specifically rotation invariant as physicians can examine a slide from different angles without tampering with the diagnostic results. In our study, we have also applied the same strategy to expand our training dataset. Each of the extracted patches  $P_{a_i}$  is rotated in the X-Y plane with  $\frac{Q\pi}{4}$  variations with  $Q$  in  $\{0, 1, 2, 3, 4, 5, 6, \text{and } 7\}$  as shown in Fig. 6.

#### 2.4. Proposed OPOD and APOD techniques

In this section, we have discussed our proposed “One Patch in One Decision” (OPOD) and “All Patches in One Decision” (APOD) techniques elaborately for performing patch-wise and image-wise classification respectively.



**Fig. 5.** (a) and (e) represent normal and in-situ carcinoma image respectively. (b), (c) and (d) depict patches extracted from the same region of (a) of sizes  $64 \times 64$ ,  $128 \times 128$  and  $512 \times 512$  respectively. (f), (g) and (h) represent patches extracted from the same region of (e) of sizes  $64 \times 64$ ,  $128 \times 128$  and  $512 \times 512$  respectively. Patches of smaller sizes as in (b), (c), (f) and (g) fail to capture any discriminating information. However, (d) and (h) gives a larger field-of-view and contains relevant diagnostic information.



**Fig. 6.** (a) Input patch, (b–i) denotes different patch rotations of the patch (a) at an angle  $\frac{Q\pi}{4}$  with  $Q$  varying from 0–7.

#### 2.4.1. One Patch in One Decision (OPOD)

One Patch in One Decision (OPOD) is mainly responsible for predicting class labels of each patch extracted from the pre-processed histopathological images. Let  $I : \Omega \rightarrow R^+$  represent a preprocessed image where  $\Omega \subset R^2 = \{(x, y) | x, y \in Z\}$  represent the image domain where  $(x, y)$  is the spatial coordinates of a pixel in an image domain  $\Omega$ . Let  $Pa_i$ , ( $i = 1, 2, 3, \dots, n$ ), be  $i^{th}$  patch extracted from  $I$  where  $n$  is the total number of patches. Let us also assume that  $T_{Pa_i}$  be the desired output for patch  $Pa_i$  used during training which can be defined as

$$T_{Pa_i} = k \text{ for } L(I) = k, \quad (2)$$

where  $k \in \{0, 1, 2, 3\}$  the set of four class labels,  $L(I)$  be the annotated class label of image  $I$

Straightforwardly the training process of the proposed patch based classifier (PBC) using OPOD technique starts with extraction of patches from the input images. The architecture for the training model is described in section 2.5. The PBC, in its training phase, automatically extracts discriminating features from each of the patches. The block diagram for patch-wise classification by OPOD technique is illustrated in Fig. 7. The preprocessed input images in our dataset are represented in (7a). Patches are extracted and augmented from these images as shown in (7b). Each of these patches is annotated with the same class label as the original image which is represented in (7c). The extracted patches and their annotated class labels are passed to the proposed PBC to train the model as depicted in (7d).

After training the PBC, the final decision of a patch class label is predicted from the scores returned by a softmax classifier. We can formulate classification by softmax classifier as follows-

Let  $M$  be a mapping function which takes as input a set of patches  $S \{Pa_1, Pa_2, Pa_3, \dots, Pa_n\}$  and maps them to their corresponding class labels by a simple dot product of the patch  $Pa_i$  and the weight matrix

$\mathcal{W}_M$ . The pixels in the patch  $Pa_i$  are flattened into a single column vector of dimension  $D \times 1$  where  $D = h \times w$  is the number of elements in the patch  $Pa_i$ ,  $h$  and  $w$  being the height and the width of the patch  $Pa_i$  respectively. The weight matrix ( $\mathcal{W}_M$ ) has dimension  $N \times D$  where the number of rows that is  $N$  represents the total number of classes and each row is a weight vector corresponding to an individual histopathological class. The mapping function  $M$  defined in equation 3 is the array of class scores or un-normalized probabilities of each class for a patch  $Pa_i$ .

$$M(Pa_i, \mathcal{W}_M) = \mathcal{W}_M \cdot Pa_i \quad (3)$$

The softmax classifier normalizes the un-normalized probabilities returned by  $M$ . The normalized class probability ( $O_l$ ) of the  $l^{th}$  class for patch  $Pa_i$  returned by the softmax classifier is given in equation 4.

$$O_l = \frac{e^{M_l}}{\sum_{k=1}^N e^{M_k}} \quad (4)$$

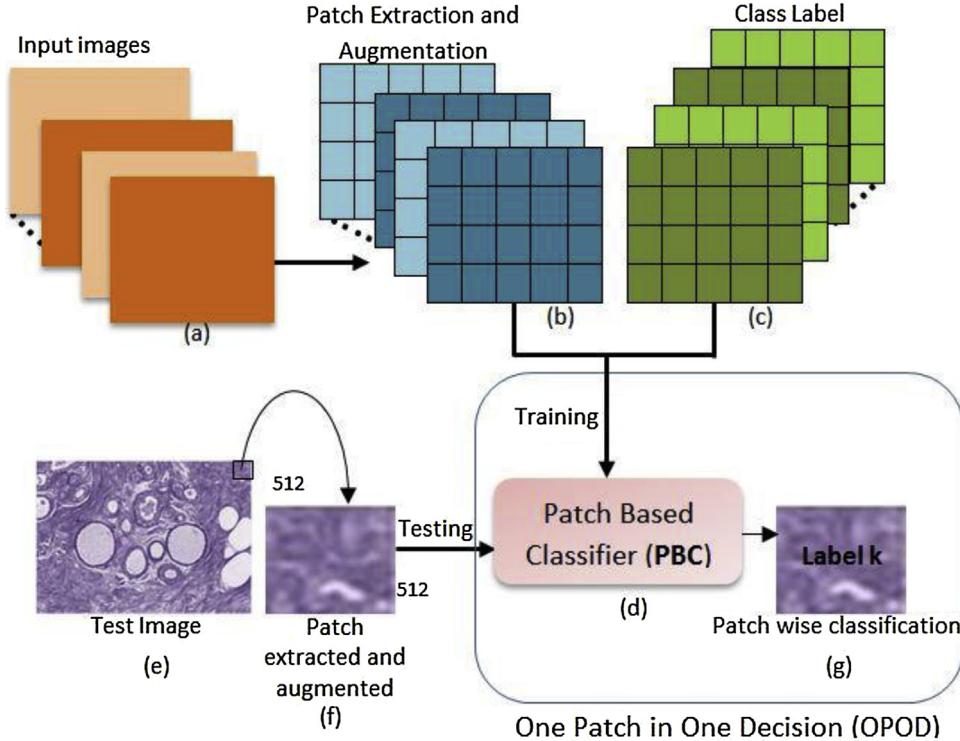
Where  $k = \{1, 2, \dots, N\}$  and  $N$  is the number of classes.

The gradient of the loss function has to be calculated to perform parameter update. The gradient calculation of the loss function is given in section 2.5.1. The score returned by the softmax classifier is updated due to a change in weight matrix in successive iterations till the loss function is minimized to a negligible value. Finally, after an unbeaten run of softmax classifier over several iterations, eventually, it will return probabilities or scores indicating that the patch belongs to one of the four predefined classes. The patch under consideration will be assigned that class label for which the corresponding probability that is the score is maximum. The class label function  $cl(Pa_i)$  predicts the class label of each input patch  $Pa_i$  according to Equation 5.

$$\gamma = cl(Pa_i) = \text{argmax}(O_l) \quad (5)$$

For example, if the softmax classifier score of class 0 for a patch  $Pa_i$  is maximum then it will be classified as class 0 patch otherwise to the other appropriate classes based on the score.

Thus, once the model is trained, it predicts the class label of each of the test patches passed to it based on the features it had learned during its training phase. The idea of predicting class labels of test patches is illustrated in Fig. 7e, 7f, and 7g. A patch (7f) extracted from a test image (7e) is passed to the trained PBC (7d) which in turn predicts the class label of this input patch as represented in 7g.



**Fig. 7.** Block Diagram of patch-wise classification by the proposed OPOD technique. (a) Input Images (b) Patch extraction and augmentation of the input images (c) Annotated class labels of the patches (d) Training of the proposed PBC (e) Test image (f) Patch extracted from the input image (g) Patch label prediction by the trained patch based classifier (PBC) where  $k \in \{0, 1, 2, 3\}$ .

If the trained PBC classifies all patches extracted from an image to the same class label as that of the image, then we say the classification to be a correct one. Depending on this criterion, the patch-wise classification accuracy, sensitivity, precision, and F1-score have been evaluated in section 3.2.

#### 2.4.2. All Patches in One Decision (AOPD)

All Patches in One Decision (AOPD) is aimed at classifying each of the input images into four different histological classes based on a majority voting scheme on its patch labels that is predicted by the proposed OPOD technique, described in the previous subsection.

Let  $Pa_i$ ,  $i = \{1, 2, \dots, n\}$ , be a patch extracted from an image  $I$  in row major order and  $\gamma$  represents the class label of patch  $Pa_i$  predicted by OPOD technique such that  $\gamma \in [0, 1, 2, 3]$  corresponding to the four different histology classes. The final class assigned to the image  $I$  is predicted by the majority voting scheme.

The counting function  $F_i^{(k)}$  can be defined as

$$F_i^{(k)} = \begin{cases} 1 & \text{if } \gamma = k \\ 0 & \text{if } \gamma \neq k \end{cases} \quad (6)$$

Where  $k \in [0, 1, 2, 3]$  corresponding to the four histology classes

The count of total votes for each class  $C^k$  can be expressed as:

$$C^k = \sum_{i=1}^n F_i^{(k)} \quad (7)$$

Say,  $T$  be the predicted class label of image  $I$ . The proposed AOPD technique assigns the image  $I$  to that histological class where its maximum patches belong and is defined as

$$T = \text{argmax}(C^k) \quad (8)$$

Thus the AOPD technique considers all the patch labels extracted from an image  $I$  to predict the class label of that image by majority

voting scheme. This idea is further illustrated by Fig. 8. The different patch labels for an input preprocessed image predicted by the OPOD technique is shown in 8a. As evident from 8a, maximum patches of the input image have class label 1 (benign class), so AOPD predicts the image label as 1 which is shown in 8b by majority voting scheme.

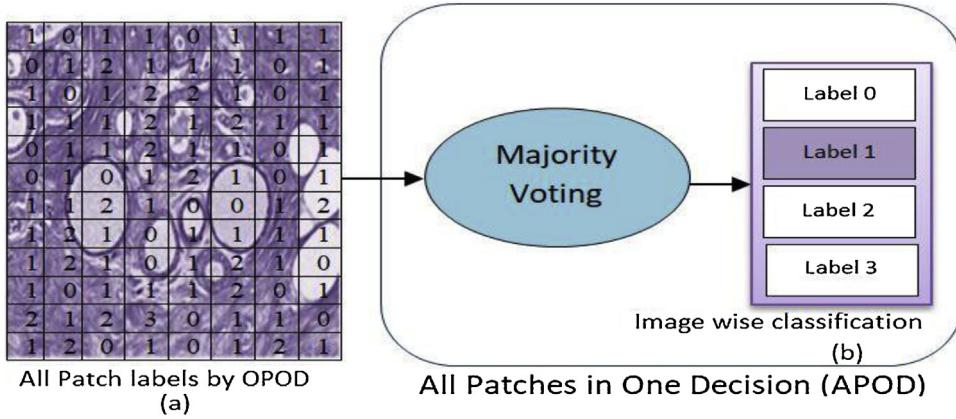
#### 2.5. Analysis of Prototypic CNN architectures

Different CNN architectures (Chen et al., 2018) vary from each other depending on the number of trainable parameters which can be obtained by varying the number of kernels in each layer, kernel size and the number of fully connected layers. In this study, we have developed various CNN architectures to study the behavioral characteristics of the model with change in network behavior (for example number of trainable parameters, network depth, number of kernels in each layer, etc.) on the dataset used in this work. Table 3 gives a vivid comparison of the different networks developed and the learning accuracy obtained in each case.

Table 3 shows that the classification results obtained using different variants of the CNN architectures on ICIAR 2018 dataset. The number of neurons in the fully connected layers has been chosen such that the number of parameters does not increase abruptly (more neurons lead to more number of parameters) since that might lead to overfitting as has been explained in the following part of this section. The last fully connected layer must have four neurons since this work is a 4-class classification problem. As evident from Table 3, maximum learning accuracy has been obtained with the 2<sup>nd</sup> model, and its details are explained in the following part of this section.

#### Proposed Convolutional Neural Network (CNN) architecture

The architecture used in this study is a hierarchical CNN where each level combines features generated from the lower levels to generate higher dimension feature maps which are used for image

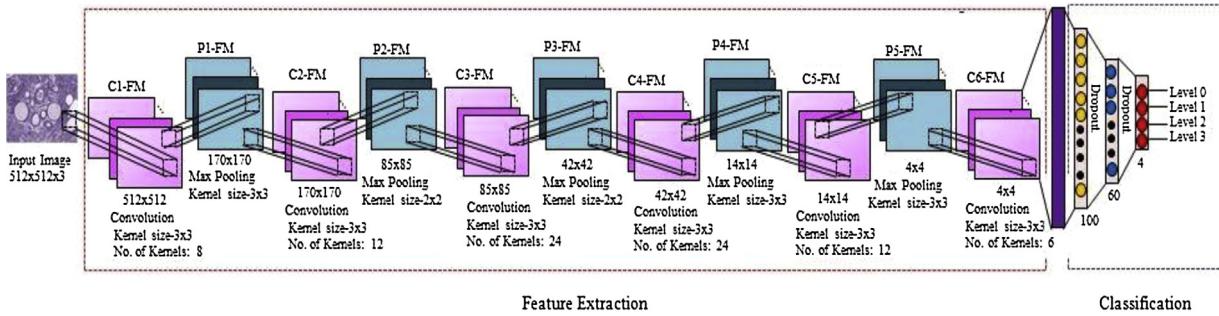


**Fig. 8.** Block Diagram of image-wise classification by the proposed APOD technique. (a) Patch labels of an image predicted by OPOD technique (b) Image label prediction based on patch label majority voting by proposed APOD technique.

**Table 3**

Different CNN architectures studied and their performance on the used dataset.

Model No.	No. of layers	No. of convolution layers	No. of convolution filters in each layer	Kernel size of convolution layers	No. of Pooling layers	Kernel Size of Pooling layers	Number of fully connected Layers	Neurons in each layer	Number of trainable parameters	Learning Accuracy (%)
1	15	6	8,12,24,24,12,6	3,3	6	(3,3),(2,2), (2,2),(2,2), (3,3),(3,3)	3	100,60,4	18,636	66.7
2	14	6	8,12,24,24,12,6	3,3	5	(3,3),(2,2), (2,2),(3,3), (3,3)	3	100,60,4	28,186	85.8
3	13	5	16,32,64,64,32	3,3	5	(3,3),(2,2), (2,2),(3,3), (3,3)	3	100,60,4	94,980	58.0
4	14	6	8,16,32,32,16,16	3,3	5	(3,3),(2,2), (2,2),(3,3), (3,3)	3	100,60,4	38,228	74.6
5	11	4	8,16,32,16	3,3	4	(3,3),(2,2), (2,2),(3,3)	3	100,60,4	3,30,560	45.5



**Fig. 9.** Block diagram of the CNN architecture used in the developed Patch-Based Classifier (PBC).

classification to its corresponding classes. The block diagram of the developed CNN architecture is shown in Fig. 9.

The detailed specification of different layers in our proposed PBC model is illustrated in Table 4.

This proposed CNN architecture successfully classifies each input patches of size  $512 \times 512$  into four different histology classes. Class 0 represents normal tissues, benign classes are represented by label 1, in situ, and invasive carcinoma are represented by labels 2 and 3 respectively. Each of the layers in the proposed PBC model is described in Fig. 9 elaborately.

- **Input layer**- This layer consists of preprocessed RGB patches from the augmented patch dataset each having size  $512 \times 512$ .
- **Convolution layer**- This is the fundamental building block of a CNN architecture consisting of multiple neurons with learnable weights (Chen et al., 2018) and is responsible for extracting discriminating features from an image and generating an output feature map. The equation for the same is given in Equation 9.

$$P(a, b) = c(a, b) * Pa_i(a, b) = \sum_{x=-u}^u \sum_{m=-y}^y c(x, m) Pa_i(a - m, b - x) \quad (9)$$

**Table 4**

Detailed specification of the different layers in the proposed PBC model and number of trainable parameters in each layer.

Layers	Type	Kernel Size	Number of Kernels (K) and Number of neurons (N)	Number of trainable parameters
0	Input	-	$3K \times 512 \times 512N$	-
1	Convolution	$3 \times 3$	$8K \times 512 \times 512N$	224
2	Max Pooling	$3 \times 3$	$8K \times 170 \times 170N$	-
3	Convolution	$3 \times 3$	$12K \times 170 \times 170N$	876
4	Max Pooling	$2 \times 2$	$12K \times 85 \times 85N$	-
5	Convolution	$3 \times 3$	$24K \times 85 \times 85N$	2616
6	Max Pooling	$2 \times 2$	$24K \times 42 \times 42N$	-
7	Convolution	$3 \times 3$	$24K \times 42 \times 42N$	5208
8	Max Pooling	$3 \times 3$	$24K \times 14 \times 14N$	-
9	Convolution	$3 \times 3$	$12K \times 14 \times 14N$	2604
10	Max Pooling	$3 \times 3$	$12K \times 4 \times 4N$	-
11	Convolution	$3 \times 3$	$6K \times 4 \times 4N$	645
12	Fully Connected ----		100N	9700
13	Fully Connected ----		60N	6060
14	Fully Connected ----		4N	224

where  $Pa_i(a, b)$  represent a pixel of patch  $Pa_i$  at spatial coordinates  $(a, b)$ ,  $P(a, b)$  is the new value of the pixel after convolution and  $c$  is the convolution kernel of size  $u \times y$ .

In this work, we have experimentally evaluated the learning accuracy obtained with a different number of convolution layers. The obtained results are depicted in Table 3. From Table 3 it can be seen that the maximum learning accuracy is obtained with six convolution layers. The initial convolution layer extracts the generic features from the input patches, so less number of filters are used in the initial layer which is then passed to the subsequent layers. As the succeeding layers are responsible for determining more complex patterns among the features extracted by the previous layers, so more number of filters are used in these layers. Thus, the deeper the network proceeds it extracts more in-depth feature information. Finally, in the converging convolution layers, less number of filters are used to find the most significant patterns from the patterns generated by its previous layers. In histopathological images, extraction of minute details like morphological structure information of an individual nucleus is significant to distinguish between different histological classes. Thus in this work, we have limited our kernel size for each layer to  $3 \times 3$  since a higher kernel size overlooks minute details.

**Activation function:** In this work, we have used hyperbolic tangent ( $\tanh$ ) activation function for the different convolution layers. This activation function has a much stronger gradient (Masci et al., 2011) than other activation functions like sigmoid which results in a more robust, easier and faster optimization of the loss function. The output (op) of a neuron at position  $z$  using  $\tanh$  activation function is represented by Equation 10 and Equation 11.

$$op = g(c * Pa_i) \quad (10)$$

where function  $g$  is defined as

$$g(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (11)$$

**Max-Pooling Layer-** This layer (Chen et al., 2018) is specialized to select the maximum activation from a non-overlapping rectangular region of size  $R_x \times R_y$ , thus down-sampling an image by a factor of  $R_x$  and  $R_y$  in an x-y direction respectively. Thus, this layer selects features which are superior and position invariant which in turn helps in faster convergence and improves generalization performance of the model. In our architecture, we have used Max Pool layers stacked with the convolution layers as in Fig. 9.

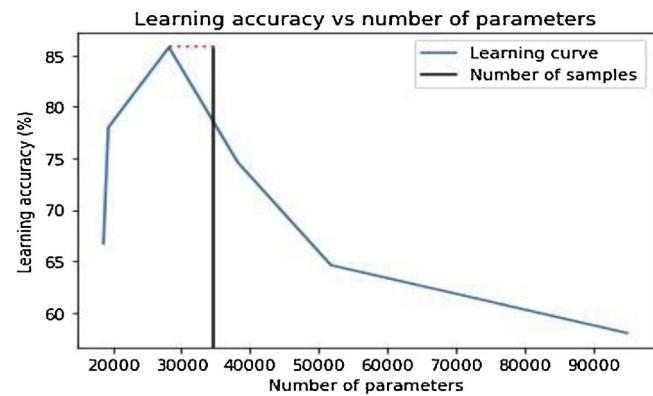
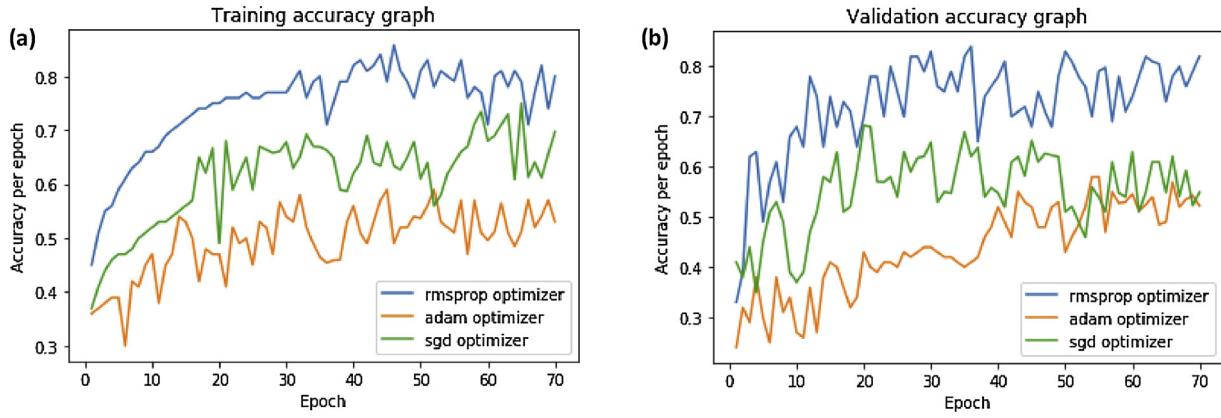


Fig. 10. Comparison of learning accuracy with the number of parameters and number of samples.

- Classification layer-** This is the output layer (Chen et al., 2018) which consists of four neurons corresponding to the 4-class classification problem (normal, benign, in-situ and invasive) which is activated by a softmax activation function.
- Dropout Layer-** Dropout layers (Srivastava et al., 2014) prevents overfitting and enhances the performance of a CNN classifier. Two dropout layers have been used between the fully connected networks to avert overfitting with probability  $p = 0.2$ .
- Kernel Selection-** The choice of kernels and their respective sizes determines the number of trainable parameters which in turn affects the performance of a CNN architecture. A study by (Overfitting In Machine Learning, 2018) suggests that if the number of feature maps or trainable parameters ( $n_{TP}$ ) is equal to or greater than the number of samples ( $s$ ), then there exists a risk of overfitting. To avoid overfitting the number of features ( $n_{TP}$ ) should be much less than the number of samples ( $s$ ) that is  $s \gg n_{TP}$ . Thus, in this work, the number of kernels in each layer and their respective sizes were chosen such that the number of trainable parameters is less than the number of samples to avoid overfitting. We have performed an experimental analysis of learning accuracy obtained with a different number of trainable parameters and the same is reported in Fig. 10. From Fig. 10, it can be observed that with an increase in the number of parameters the learning accuracy of the system increases initially. However, on further increasing the number of trainable parameters the accuracy of the system dropped drastically. The black line denotes our number of samples, and the red dotted line significantly shows the highest learning accuracy obtained for the number of samples in comparison to the number of trainable parameters.



**Fig. 11.** (a): Comparison of training accuracy obtained by different optimizers (b): Comparison of validation accuracy obtained by different optimizers.

### 2.5.1. Learning Technique of the proposed PBC

The CNN learns by measuring the loss function which is an indicator of the error incurred in the learned trainable parameters. The objective of the CNN is to compute the trainable parameters in a way such that the loss function is minimized. In this work, the categorical cross entropy loss function (Litjens et al., 2017) has been used, given in equation 12.

$$E = - \sum_{l=1}^N T_l \log O_l \quad (12)$$

where  $T_l$  and  $O_l$  represent the true probability and predicted probability by the softmax classifier for class  $l$  respectively and  $N$  is the total number of classes.

The symbols  $O$  and  $M$  have the same meanings as defined in section 2.4.1.

Next, the gradient of the loss function has to be derived to perform the parameter or weight ( $M$ ) update.

$$\frac{\partial E}{\partial M_i} = - \sum_{l=1}^N T_l \frac{\partial \log O_l}{\partial M_i} \quad (13)$$

$$\frac{\partial E}{\partial M_i} = - \sum_{l=1}^N T_l \frac{\partial \log O_l}{\partial O_l} \frac{\partial O_l}{\partial M_i} \quad (14)$$

$$\frac{\partial E}{\partial M_i} = - \sum_{l=1}^N \frac{T_l}{O_l} \frac{\partial O_l}{\partial M_i} \quad (15)$$

Assume that we know the value of  $\frac{\partial O_l}{\partial M_i}$  for the time-being (proof given in appendix 1), that is

$$\frac{\partial O_l}{\partial M_i} = \begin{cases} O_l(1 - O_l) & l = i \\ -O_l O_i & l \neq i \end{cases} \quad (16)$$

where  $i$  is the correct class.

We put the value of  $\frac{\partial O_l}{\partial M_i}$  in equation 15 and get equation 17. Equation 17 is divided into two parts, one for  $l = i$  and the other for  $l \neq i$ .

$$\frac{\partial E}{\partial M_i} = - \frac{T_i}{O_i} \times O_i(1 - O_i) - \sum_{l \neq i} \frac{T_l}{O_l} (-O_l O_i) \quad (17)$$

The first part of Equation 17,  $(-\frac{T_i}{O_i} \times O_i(1 - O_i))$  represents the condition when  $l = i$  and the second part that is  $(\sum_{l \neq i} \frac{T_l}{O_l} (-O_l O_i))$  represents the scenario when  $l \neq i$ .

$$\frac{\partial E}{\partial M_i} = -T_i(1 - O_i) + \sum_{l \neq i} T_l O_i \quad (18)$$

$$\frac{\partial E}{\partial M_i} = -T_i + T_i O_i + \sum_{l \neq i} T_l O_i \quad (19)$$

$$\frac{\partial E}{\partial M_i} = -T_i + O_i \left( \sum_{l=1}^N T_l \right) \quad (20)$$

Now  $\sum_{l=1}^N T_l = 1$ , because the sum of all the probabilities in a distribution is 1. Thus equation 20 can be further reduced to equation 21.

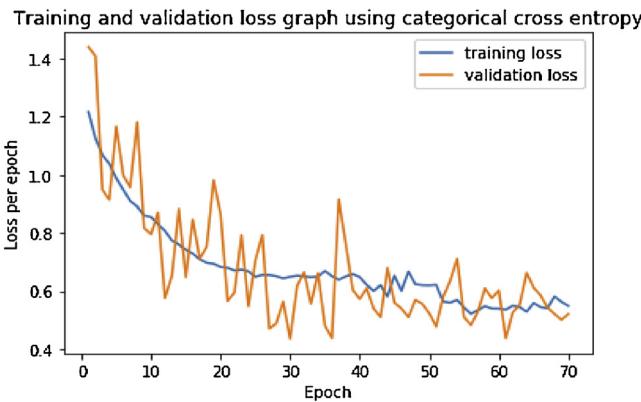
$$\frac{\partial E}{\partial M_i} = -T_i + O_i \quad (21)$$

Finally, equation 21 will be used to perform the score update in the next iteration and the process will be repeated in successive iterations till the loss function is minimized to a very small value. To accelerate the training process, root mean square propagation (rmsprop) optimizer (Lecture 6.5- RMSprop, 2018) has been used with tuned hyperparameters. The learning rate has been experimentally set to 0.001, and all the weights in this network are initialized randomly. A total of 70 epochs has been used in this work. Fig. 11(a) represents the training accuracy, and Fig. 11(b) represents the validation accuracy obtained at different epochs using three different optimizers namely rmsprop (Lecture 6.5- RMSprop, 2018), adam optimizer (Kingma and Jimmy, 2014), and sgd (Sutskever et al., 2013). As evident from the figure, the maximum accuracy for both training and validation is reported for the rmsprop optimizer.

We have also shown the variation of the training and validation loss for each epoch incurred using categorical cross entropy loss function. The same is reported in Fig. 12.

### 3. Experimental Results and Discussion

In this section, we evaluate the classification performance of our proposed model in terms of sensitivity, precision, F1-score, and accuracy. We have initially split the training set into three parts for training, validation, and test (details in section 3.1.) and



**Fig. 12.** Training and validation loss per epoch curve.

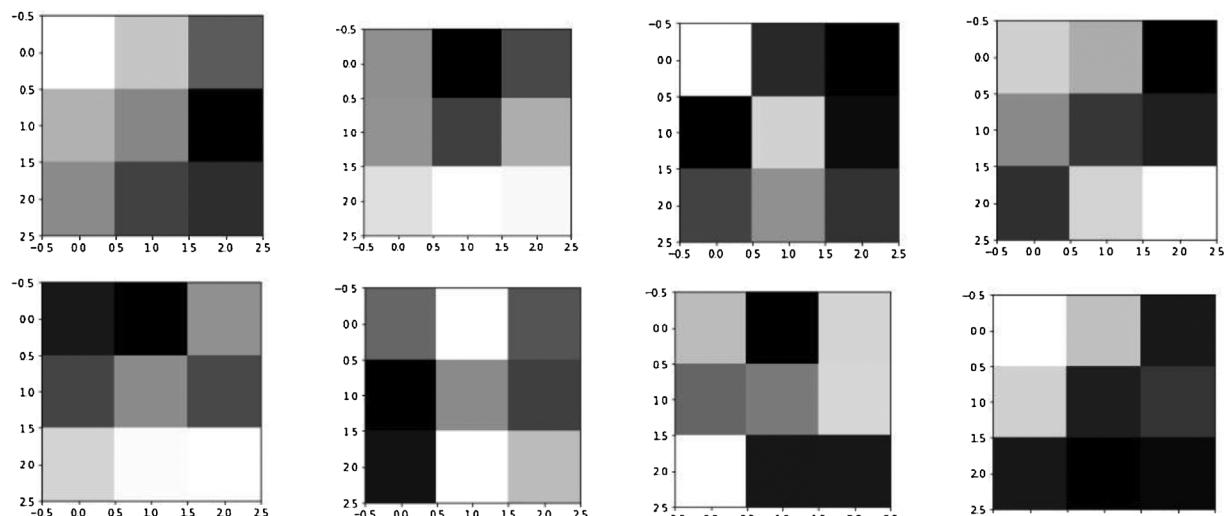
have reported the classification performance of both patch-wise and image-wise classification on this test set. In addition, the accuracy obtained in classifying the histology images in the hidden test dataset of the challenge has also been reported.

Our proposed model is implemented in python using TensorFlow and Keras library on a CPU based system with Intel Xeon Processor, 128 GB RAM.

### 3.1. Dataset Description

The dataset used in this work is ICIAR 2018 challenge dataset ([ICIAr, 2018](#)) which consists of 400 Hematoxylin and eosin (H&E) stained breast histology microscopic images for training purpose and a separate hidden test set consisting of 100 images. The training dataset has even distribution of images in each class. The training set images have been annotated into four different categories namely normal, benign, in-situ and invasive carcinoma. Two medical experts performed the annotations of these images. Each of the images has  $2048 \times 1536$  pixels with pixel size  $0.42 \mu\text{m} \times 0.42 \mu\text{m}$ . However, the class labels of the hidden test dataset were not disclosed.

As a standard practice, the training set was split into three parts for training, validation, and testing. 10 images from each histological classes in the training dataset were separated from reporting the image-wise classification performance of our model, reported in section 3.3. The remaining 360 images were used to develop the proposed model as shown in [Table 5](#).



**Fig. 13.** Weights learned by the first convolutional layer.

**Table 5**  
the Image-wise split of the ICIAR-2018 dataset.

Type	Training	Validation	Testing
Normal	70	20	10
Benign	70	20	10
In-situ	70	20	10
Invasive	70	20	10
<b>Total</b>	<b>280</b>	<b>80</b>	<b>40</b>

**Table 6**  
the Patch-wise split of the ICIAR-2018 dataset.

	Training (70%)	Validation (10%)	Testing (20%)
Number of patches	24192	3456	6912

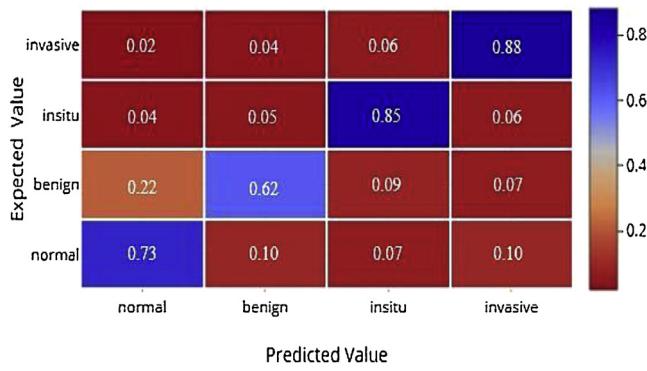
Patch extraction and augmentation were performed on the 360 images resulting in a total of 34,560 patches. The details of which is described in section 2.3. 10% of 34,560 samples that is 3456 samples were used for validation purpose, 20% of 34,560 samples that is 6912 samples were used for testing to report the classification accuracy for patch-wise classification, details in section 3.2 and the remaining number of samples that is 24,192 samples were used for training to develop the model as shown in [Table 6](#).

Also, we have evaluated the performance of our model on the hidden test set of the ICIAR-2018 breast histology challenge which is described elaborately in section 3.4

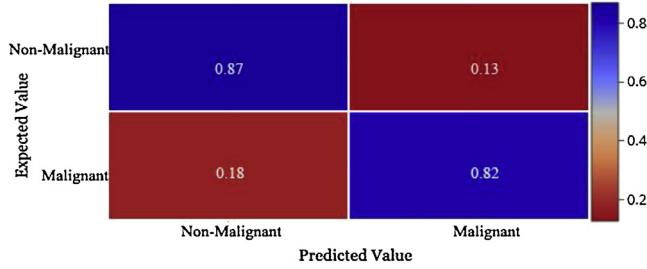
### 3.2. Patch-wise Classification by OPOD

After training and validating the developed model on 27,648 samples (24,192 + 3,456), 6,912 patches were used for testing to measure the patchwise classification accuracy. Our trained patch based classifier (PBC) assigned a class label to each of the test patches. The patches were labelled as either class 0 denoted by [1,0,0,0], class 1 by [0,1,0,0], class 2 by [0,0,1,0] or class 3 by ([Rangayyan et al., 2007](#)) for 4-histopathological classes. For 2-class classification, class 0 was denoted by [1,0] and class 1 was denoted by ([Rangayyan et al., 2007](#)). For 2-class classification benign and normal images were considered in the same group that is non-malignant and in-situ, invasive carcinoma was clubbed in the second group that is malignant class.

The average patch-wise classification sensitivity achieved by our proposed classifier on the test set (obtained by splitting training



**Fig. 15.** Normalized Heat map for patch-label prediction (4-class).



**Fig. 16.** Normalized Heat map generated for patch-label prediction (2-class).

**Table 7**  
Patchwise classification by OPOD for 4 class and 2 class.

Class	Normal	Benign	In-situ	Invasive	Average
Sensitivity (%)	74	62	85	88.60	77.40
Precision (%)	71.4	77.37	80.68	80.30	77.44
F1-Score (%)	72	68.80	82.70	84.26	76.94
Class	Non-Malignant	Malignant			
Sensitivity (%)	87		82		84.50
Precision (%)	84.72		84.88		84.80
F1-Score (%)	85.90		83.50		84.70

set) for 4-class classification is 77.4 %, and for 2-class classification, the reported sensitivity was 84.5%.

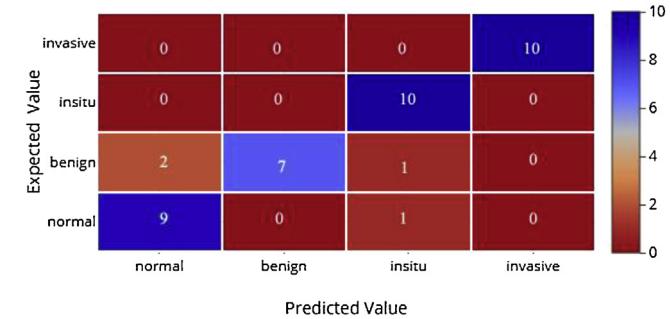
The weight learned by the eight filters in convolution layer 1 is given in Fig. 13. The filtered images obtained corresponding to the eight filters after the first layer of convolution is presented in Fig. 14. The heat map of the confusion matrix for 4-class and 2-class classification are shown in Figs. 15 and 16 respectively. To generate the heat map the values of the confusion matrix has been normalized with all values lying between (Rangayyan et al., 2007).

Table 7 illustrates the sensitivity, precision, and f1-score obtained for four class and two class patch-wise classification. As evident from Table 7, the average sensitivity, specificity is more for 2-class classification than those of 4-class classification.

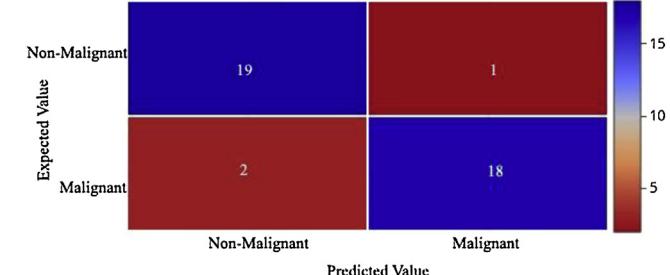
The overall accuracies reported by our patch-based classifier (PBC) for patch-wise classification using the OPOD technique for 4-class classification and 2-class classification are 77.4% and 84.7% respectively. For 4-class classification, the sensitivity achieved by the developed classifier was high for class 2(in-situ) and class 3(invasive) but low for Class 0 (Normal class) and Class 1 (Benign Class) because of the immense similar features amongst these two classes. As a result, many patches of class 0 were labeled as class 1 and vice versa. However, when two breast-histopathological classes were considered, then the sensitivity increased because in this case normal and benign images were considered in the same category that is non-malignant.

**Table 8**  
Image wise classification by APOD for 4 class and 2 class.

Class #	Normal	Benign	Insitu	Invasive	Average
Sensitivity (%)	90	70	100	100	90
Precision (%)	75	100	90.90	100	91.48
F1-Score (%)	81.80	82.35	95.23	100	89.85
Class #	Non-Malignant	Malignant			
Sensitivity (%)	95		90		92.50
Precision (%)	90.47		94.70		92.59
F1-Score (%)	92.68		92.30		92.49



**Fig. 17.** Heat map for image-label prediction (4-class).



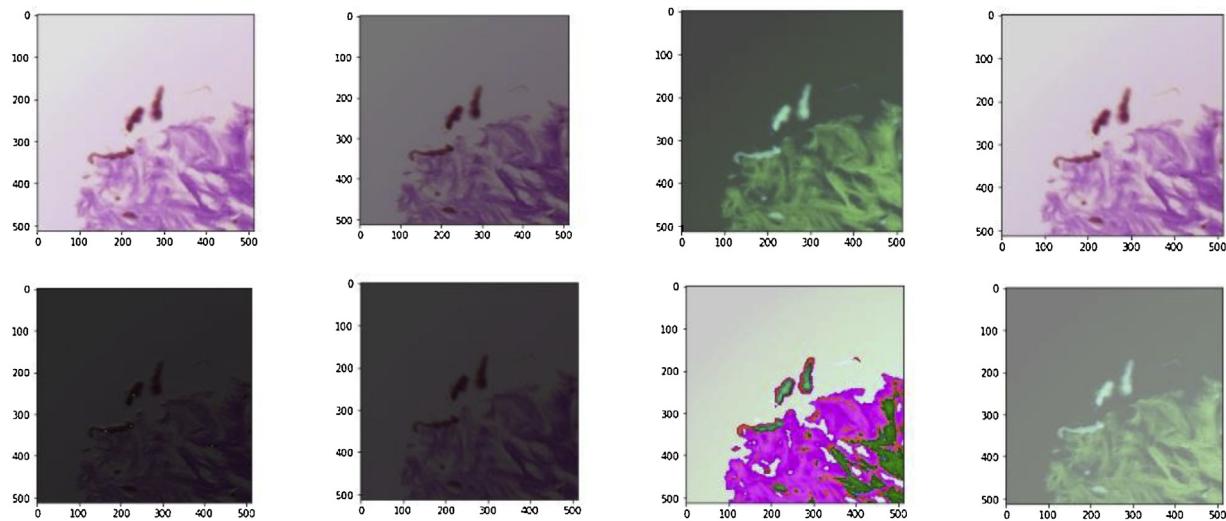
**Fig. 18.** Heat map for image-label prediction (2-class).

### 3.3. Image-wise classification by APOD

As mentioned in section 3.1, for 4-class classification, 10 images from each class were separated for predicting the image-wise classification accuracy by APOD technique. However, for 2-class classification 20 images were considered in non-malignant category and 20 in the malignant category. This technique predicted the class labels of each image from its extracted patch labels (predicted by the proposed PBC) using majority voting. Table 8 below shows the sensitivity, precision, and f1-score obtained for image-wise prediction (obtained by splitting training set) for 4-class classification and 2-class classification respectively.

The overall accuracies of the APOD technique for image-wise label prediction for 4-class classification and 2-class classification are 90% and 92.5% respectively. The heat map of the confusion matrix for image-wise label prediction for 4-class and 2-class classification is given in Figs. 17 and 18 respectively.

As seen from Table 7 and Table 8, the sensitivity, precision, and F1-score obtained by APOD technique are better than OPOD for both 4-class and 2-class classification. This is because, in OPOD, the decision about the class labels is made by considering each patch separately. In contrary, APOD technique considers all the patches as a unit and assigns the image to that class where its maximum patches belong ignoring many misclassified patches. As a result, APOD performs better than OPOD for 4-class and 2-class classification.



**Fig. 14.** Filtered images obtained after the first convolution layer.

**Table 9**

Comparative analysis of the performance of the proposed PBC for 4-class and 2-class histopathological image classification.

Method	Database Used	Number of histopathological classes	Accuracy (Patch-wise classification)	Accuracy (Image-wise classification)
Vang et. al. (2018)	ICCIAR 2018	4 (normal, benign, in-situ, invasive)	Not Provided	87.5%
Our proposed PBC	ICCIAR 2018	4(normal, benign, in-situ, invasive)	77.4%	90%
Kaymak et.al. (2017)	Private Dataset	2 (cancerous, non-cancerous)	Not Provided	70.4%
Cruz-Roa et.al. (2014)	Private dataset	2 (invasive, non-invasive)	79.6%	Not Provided
Spanhol et.al. (2016)	Breakhis dataset	2 (benign and malignant)	84.6%	Not Provided
Our proposed PBC	ICCIAR 2018	2(non-malignant and malignant)	84.7%	92.5%

### 3.4. Image-wise classification on the hidden test set

In this section, we discuss the performance of our model on the hidden test dataset of ICIAR-2018 challenge consisting of 100 images. The class labels of this set are not known to us, and hence the task of this challenge was to classify each of the hidden test data into its four histopathological classes. Our model achieved an accuracy of 87% on this hidden test dataset. However, we could not provide the class-wise sensitivity, precision, and F1-score for this part since we had only access to the accuracy of the hidden test set.

### 3.5. Comparison with the existing state of the art

This section is subdivided into two sub-sections. The first sub-section (3.5.1) performs a comparative analysis of our proposed model with other state-of-the-art methods for histopathological image classification in terms of patch wise and image wise classification (resulting from training set split) as shown in Table 9. The next subsection (3.5.2) compares the accuracy obtained by our model in classifying the hidden test data of ICIAR-2018 challenge with the other state-of-the-art methods of that challenge in Table 10.

#### 3.5.1. Patch-wise and Image-wise classification resulting from training set split

In this section, we elaborately compare the performance of our model in terms of patch wise and image wise classification on the test set resulting from training dataset split as shown in Table 9. For the sake of fair comparison, we have compared our model with (Vang et al., 2018) since they use the same dataset as ours and perform a similar train-validation-test split on the training set. However, in (Vang et al., 2018) the accuracy obtained on classifying the hidden test set of ICIAR-2018 challenge is not reported.

The other methods in this section were included to compare the performance of our model with the other state of the art for 2-class classification.

Further (Vang et al., 2018) does not evaluate the performance of their system for patch-wise classification, but in this work, we have evaluated the performance of the classifier for both patch-wise and image-wise classification.

It can be seen in Table 9, that the lowest accuracy for 2-class classification is reported (Kaymak et al., 2017). This is evident from the fact that (Kaymak et al., 2017) uses handcrafted features for classification which relies on the accuracy of the nuclei extraction phase since the structures and morphology of the nuclei are evaluated to determine its histopathological class. However, the performance

**Table 10**

Comparative analysis with the ICIAR-2018 state-of-the-art methods.

Method	Accuracy	Method	Accuracy	Method	Accuracy
Chennamsetty et al.	87%	Nedjar et al.	81%	Ferreira et al.	76%
Kwok	87%	Ravi et al.	80%	Pimkin et al.	76%
<b>Our Proposed model</b>	<b>87%</b>	Wang et al.	79%	Sarker et al	75%
Brancati et al.	86%	Cao et al.	79%	Rakhlin et al.	74%
Marami et al.	84%	Seo et al.	79%	Iesmantas et al	72%
Kohl et al.	83%	Sidhom et al.	78%	Xie et al.	72%
Wang et al.	83%	Guo et al.	77%	Weiss et al	72%
Steinfeldt et al.	81%	Ranjan et al.	77%	Awam et al	71%
Kone et al.	81%	Mahbod et al.	77%	Liang	70%

of deep learning based frameworks are not limited by the classification results of nuclei segmentation step since training and classification using deep learning is based on the direct processing of image regions. Our work handles a much complex dataset than (Fabio Alexandre Spanhol et al., 2016) and (Cruz-Roa et al., 2014) since the ICIAR dataset has images of four different varieties (normal, benign, in-situ, invasive), unlike others where images are either normal or malignant.

### 3.5.2. Comparison with ICIAR-2018 state-of-the-art methods

In this section, we compare the accuracy obtained by our model in classifying the images in the ICIAR-2018 hidden test dataset with methods reported in (Aresta et al., 2018) as shown in Table 10.

As evident from Table 10, the highest accuracy obtained in (Aresta et al., 2018) is 87%. Our method also achieves accuracy of 87% for classifying images in ICIAR-2018 test set.

## 4. Conclusion

In this work, we have developed a patch-based classifier (PBC) for automatic classification of ICIAR-2018 breast histology dataset into four class namely normal, benign, in situ and invasive carcinoma. The number of filters in each layer, kernel size was adjusted in such a way that the number of trainable parameters is less than the number of samples so as to prevent overfitting. This proposed classifier first predicts the class label of each input patch by OPOD technique. The whole image label was finally predicted by APOD technique. This classifier achieves an accuracy of 87% on the hidden test set of the challenge. As a part of our future research, we will try to implement the proposed model on another histopathological image dataset.

## Declarations of interest

None.

## Acknowledgments

The first author is grateful to the Department of Science and Technology (DST), Government of India for providing her Junior Research Fellowship (JRF) under DST INSPIRE fellowship program (IF170366).

## Appendix A

The derivation of  $\frac{\partial O_l}{\partial M_i}$  is shown in this section. We know from section 2.4.1 that the predicted probability of softmax classifier is given as-

$$O_l = \frac{e^{M_l}}{\sum_{k=1}^N e^{M_k}} \quad (22)$$

Since  $l$  can take any class from  $\{1, 2, \dots, N\}$ , we divide it into two sets. One set will represent the desired class that is  $\{l = i\}$ , and the other set represent all the other classes excluding  $\{i\}$  that is  $l \neq i$  and  $N$  is the total number of classes.

Case 1:

1)  $l = i$

In this case  $\frac{\partial O_l}{\partial M_i}$  will be derived by quotient rule as follows:-

$$\frac{\partial O_l}{\partial M_i} = \frac{\left(\frac{\partial(e^{M_l})}{\partial M_i} \times \sum e^{M_k}\right) - \left(\frac{\partial\left(\sum e^{M_k}\right)}{\partial M_i} \times e^{M_l}\right)}{\left(\sum e^{M_k}\right)^2} \quad (23)$$

In this case,  $\frac{\partial(e^{M_l})}{\partial M_i} = e^{M_l} = e^{M_i}$  (since  $l = i$ ) and  $\frac{\partial\left(\sum e^{M_k}\right)}{\partial M_i} = e^{M_i}$  due to the reason given below:

$$\begin{aligned} \frac{\partial\left(\sum e^{M_i}\right)}{\partial M_i} &= \frac{\partial(e^{M_1})}{\partial M_i} + \frac{\partial(e^{M_2})}{\partial M_i} + \dots \\ &+ \frac{\partial(e^{M_i})}{\partial M_i} + \dots + \frac{\partial(e^{M_N})}{\partial M_i} = 0 + 0 + \dots + e^{M_i} + \dots + 0 = e^{M_i} \end{aligned} \quad (24)$$

$$\frac{\partial O_l}{\partial M_i} = \frac{(e^{M_l} \times \sum e^{M_k}) - (e^{M_i} \times e^{M_l})}{\left(\sum e^{M_k}\right)^2} \quad (25)$$

$$\frac{\partial O_l}{\partial M_i} = \frac{e^{M_l}(\sum e^{M_k} - e^{M_i})}{\left(\sum e^{M_k}\right)^2} \quad (26)$$

$$\frac{\partial O_l}{\partial M_i} = \frac{e^{M_l}}{\sum e^{M_k}} \cdot \frac{(\sum e^{M_k} - e^{M_i})}{\sum e^{M_k}} \quad (27)$$

$$\frac{\partial O_l}{\partial M_i} = O_l(1 - \frac{e^{M_i}}{\sum e^{M_k}}) \quad (28)$$

$$\frac{\partial O_l}{\partial M_i} = O_l(1 - O_i) \quad (29)$$

Case 2:

2)  $l \neq i$

Again, in this case, we will apply the quotient rule to solve the derivation of  $\frac{\partial O_l}{\partial M_i}$ .

$$\frac{\partial O_l}{\partial M_i} = \frac{\left(\frac{\partial(e^{M_l})}{\partial M_i} \times \sum e^{M_k}\right) - \left(\frac{\partial\left(\sum e^{M_k}\right)}{\partial M_i} \times e^{M_l}\right)}{\left(\sum e^{M_k}\right)^2} \quad (30)$$

In this case,  $\frac{\partial(e^{M_l})}{\partial M_i} = 0$  (since  $l \neq i$ )

$$\frac{\partial O_l}{\partial M_i} = \frac{(0 \times \sum e^{M_l}) - (e^{M_i} \times e^{M_l})}{(\sum e^{M_k})^2} \quad (31)$$

$$\frac{\partial O_l}{\partial M_i} = \frac{-e^{M_i} e^{M_l}}{(\sum e^{M_k})^2} = \frac{-e^{M_i}}{\sum e^{M_k}} \frac{e^{M_l}}{\sum e^{M_k}} \quad (32)$$

$$\frac{\partial O_l}{\partial M_i} = -O_i O_l \quad (33)$$

The above two cases, Case 1 and Case 2 (equation 29 and equation 33) can be written as in Equation 34.

$$\frac{\partial O_l}{\partial M_i} = \begin{cases} O_l(1 - O_l) & l = i \\ -O_i O_l & l \neq i \end{cases} \quad (34)$$

## References

- Rangaraj, Rangaraj M., et al., 2007. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute* 344 (3–4), 312–348.
- World Health Organization, 2018. (WHO) report on "Cancer" (accessed 18 January 2018) [www.who.int/mediacentre/factsheets/fs297/en/](http://www.who.int/mediacentre/factsheets/fs297/en/).
- Surakasula, Aruna, et al., 2014. A comparative study of pre-and post-menopausal breast cancer: Risk factors, presentation, characteristics and management. *Journal of research in pharmacy practice* 3 (1), 12–18.
- Behrens, Sarah, et al., 2007. Computer assistance for MR based diagnosis of breast cancer: present and future challenges. *Computerized medical imaging and graphics* 31 (4–5), 236–247.
- He, Lei, et al., 2012. Histology image analysis for carcinoma detection and grading. *Computer methods and programs in biomedicine* 107 (3), 538–556.
- Veta, Mitko, et al., 2013. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. *PloS one* (8), 7.
- Jain, Aashna, et al., 2014. Cancerous cell detection using histopathological image analysis. *International Journal of Innovative Research in Computer and Communication Engineering* 2 (12), 7419–7426.
- Basavanhally, Ajay, et al., 2013. Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides. *IEEE Transactions on biomedical engineering* 60 (8), 2089–2099.
- Breast cancer and breast pathology, <http://pathology.jhu.edu/breast/grade.php>, (accessed 10 January 2018).
- Khan, Adnan M., et al., 2013. HyMap: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. *Journal of pathology informatics* 4 (Suppl).
- MITOS Dataset, 2018 [http://ludo17.free.fr/mitos\\_2012/dataset.html](http://ludo17.free.fr/mitos_2012/dataset.html), (accessed 18 January 2018).
- Roullier, Vincent, et al., 2016. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics* 35 (7–8), 603–615.
- Kaymak, Sertan, Helwan, Abdulkader, Uzun, Dilber, 2017. Breast cancer image classification using artificial neural networks. *Procedia Computer Science* 120, 126–131.
- Fabio Alexandre Spanhol, et al., 2016. Breast cancer histopathological image classification using convolutional neural networks. *IEEE International Conference on Neural Networks (IJCNN)*, pp. 2560–2567.
- Cruz-Roa, Angel, et al., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: SPIE Proceedings on Medical Imaging: Digital Pathology., pp. 9041.
- Wahab, Noorul, et al., 2017. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. *Computers in biology and medicine* 85, 86–97.
- Vang, Yeeleung S., Chen, Zhen, Xiaohui, Xie., arXiv preprint arXiv:1802.00931 2018. Deep Learning Framework for Multi-class Breast Cancer Histology Image Classification.
- Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, Wojna, Zbigniew, 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.
- Azevedo Tosta, Tháfina A., et al., 2017. Segmentation methods of H&E-stained histological images of lymphoma: A review. *Informatics in Medicine Unlocked* 9, 35–43.
- Macenko, Marc, et al., 2009. A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, ISBI'09, 1107–1110.
- Khan, Adnan Mujahid, et al., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering* 61 (6), 1729–1738.
- Deng, Li, Dong, Yu., 2014. Deep learning: methods and applications. *Foundations and Trends in Signal Processing* 7 (3–4), 197–387.
- Irshad, Humayun, Veillard, Antoine, Roux, Ludovic, Racoceanu, Daniel, 2014. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering* 7, 97–114.
- ICIR, 2018. Grand challenge on breast cancer histology images (accessed 11 March 2018) <https://iciar2018-challenge.grand-challenge.org>.
- Krizhevsky, Alex, et al., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.
- Chen, Liang-Chieh, et al., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE transactions on pattern analysis and machine intelligence* 40 (4), 834–848.
- Masci, Jonathan, et al., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In: *Springer International Conference on Artificial Neural Networks*, pp. 52–59.
- Srivastava, Nitish, et al., 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15 (1), 1929–1958.
- Overfitting In Machine Learning, <https://prateekjoshi.com/2013/06/09/overfitting-in-machine-learning/>, (accessed 11 March 2018).
- Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, van der Laak, Jeroen AWM, van Ginneken, Bram, Sanchez, Clara I., 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42, 60–88.
- Lecture 6.5- RMSprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, <https://www.coursera.org/lecture/neural-networks/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude-YQHk>, (accessed 15 March 2018).
- Kingma, Diederik P., Jimmy, Ba., arXiv preprint arXiv:1412.6980 2014. Adam: A method for stochastic optimization.
- Sutskever, Ilya, Martens, James, Dahl, George, Hinton, Geoffrey, 2013. On the importance of initialization and momentum in deep learning. In: *International conference on machine learning*, pp. 1139–1147.
- Aresta, Guilherme, Araújo, Teresa, Kwok, Scotty, Chennamsetty, Sai Saketh, Safwan, Mohammed, Alex, Varghese, Marami, Bahram, et al., arXiv preprint arXiv:1808.04277 2018. BACH: Grand Challenge on Breast Cancer Histology Images.