This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access



# A Deep Learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images

Nadia Brancati, Giuseppe De Pietro, *Member, IEEE*, Maria Frucci, Daniel Riccio, *Member, IEEE*

***Abstract*— Accurately identifying and categorizing cancer structures/sub-types in histological images is an important clinical task involving a considerable workload and a specific subspecialty of pathologists. Digitizing pathology is a current trend that provides large amounts of visual data allowing a faster and more precise diagnosis through the development of automatic image analysis techniques. Recent studies have shown promising results for the automatic analysis of cancer tissue by using deep learning strategies that automatically extract and organize the discriminative information from the data. This paper explores deep learning methods for the automatic analysis of Hematoxylin and Eosin stained histological images of breast cancer and lymphoma. In particular, a deep learning approach is proposed for two different use cases: the detection of invasive ductal carcinoma in breast histological images and the classification of lymphoma sub-types. Both use cases have been addressed by adopting a Residual Convolutional Neural Network which is part of a Convolutional Autoencoder Network (i.e. FusionNet). The performances have been evaluated on public datasets of digital histological images and have been compared with those obtained by using different deep neural networks (UNet and ResNet). Additionally, comparisons with the state of the art have been considered, in accordance with different deep learning approaches. The experimental results show an improvement of $5.06\%$ in F-measure score for the detection task, and an improvement of $1.09\%$ in the accuracy measure for the classification task.**

***Index Terms*— histological images, deep learning, multi-classification, detection**

## I. Introduction

The digitalization of histological specimens by using modern whole-slide digital scanners brings not only the advantage of an easy storage, visualization, and analysis of the images, but also affords the possibility of applying automatic image analysis techniques to digital histological slides to provide accurate quantifications (e.g., tumor extent and nuclei counts) and classifications of tumor sub-types with the aim both of reducing inter- and intra-reader variability among pathologists and of accelerating the diagnosis process. Any automatic analysis of digital histological images is a very challenging task, since both the spatial arrangement of the structures, e.g. nuclei and stroma, and the color distribution can be very different, also for images belonging to the same tumor class. Deep learning (DL) approaches are particularly suitable to address these problems and to perform tasks, such as the detection of specific areas of the disease or the discrimination between the tumor classes of interest. Indeed, especially when a large number of samples are available for training, a DL system learns representative features automatically and directly from the digital images of tumor tissue, with the goal of obtaining a maximum separability between the classes of structures or tumor sub-types.

In such a context, this paper presents a DL approach addressing two different use cases: i) the detection of invasive ductal carcinoma (IDC) of breast cancer, and ii) the lymphoma multi-classification in chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and matle cell lymphoma (MCL).

IDC is the most common form of invasive breast cancer and its precise detection on whole-slide images (WSI) is crucial to the diagnosis and sub-sequent estimation of grading the tumor aggressiveness of breast cancer. Manual IDC detection is tedious and time-consuming for pathologists and could be influenced by significant inter- and intra-pathologist variability in the diagnosis and interpretation of specimens. The diagnosis of lymphoma is a problematic and difficult process for pathologists. Lymphoma is a type of cancer affecting the lymphatic system and it is classified in different sub-types. In particular, three of these sub-types, CLL, FL and MCL account for $70\%$ of lymphoma cases. The most important diagnostic criterium for lymphoma are the morphological features of the tumor which can interpreted by an experienced hematopathologist, in such a way as to make a further differentiation between malignancy types to guide treatment decisions.

Both detection and classification tasks increase the demand for a reduction of workloads and of inter- and intra-observer variability, and also imply sub-specialty requirements in pathology. As a result, there is a great interest in DL networks which have the potential to reduce these workloads and augment the diagnostic capabilities of pathologists. The choice of a DL network and the training strategy to apply for a given task, depend on the type of pathological analysis to be performed.

N. Brancati, G. De Pietro, and M. Frucci are with Institute for High Performance Computing and Networking National Research Council of Italy (ICAR-CNR), Naples, Italy. e-mail: (nadia.brancati, giuseppe.depietro, maria.frucci)@cnr.it.

D. Riccio is with Institute for High Performance Computing and Networking National Research Council of Italy (ICAR-CNR), Naples, Italy and with Universita' Federico II, Naples, Italy. e-mail: daniel.riccio@unina.it

In this paper, both use cases are addressed by adopting the residual convolutional autoencoder FusionNet, in relation to two different scenarios:

- the convolutional autoencoder FusionNet is trained under a sparsity constraint in an unsupervised manner; and
- a residual convolutional neural network that is the encoding part of FusionNet is trained in a supervised manner.

In both scenarios, the features learned at the end of the encoding stage are used for the classification, by means of a softmax classifier. Indeed, also the IDC detection is addressed as a classification task, by splitting the WSI into patches and classifying each patch as IDC or non-IDC.

The perfomances of each scenario have been evaluated on public datasets [2], containing histological images acquired by using Hematoxylin and Eosin (H&E) staining technique and in relation to selected pathological use cases. We also investigated the performances of different DL networks (ResNet and UNet) for both use cases. Moreover, comparisons with different approaches in literature, working on the same datasets, have been taken into account. The performance evaluation of these methods has been given in terms of F-measure score and Balanced Accuracy (BAC) for the IDC detection and in terms of Accuracy for the lymphoma multi-classification. On observing the classification performance using overall validation and test accuracies, our second scenario has produced favorable results for both use cases. Moreover, our results outperform the state of the art. Our approach achieved the best quantitative results both for IDC detection (F-measure and BAC equal to $81.54\%$ and $87.76\%$, respectively) and for lymphoma multi-classification (Accuracy equal to $97.67\%$).

The rest of the paper is organized as follows: previous related works are presented in Section II; a description of the approach is presented in Section III; the experimental setup, comparative strategies and results for the two different use cases are discussed in Section IV; and finally, in Section V certain conclusions are drawn.

## II. Related Work

A large number of papers have been published concerning the detection and classification of histological images. Some papers propose methods that use various image processing and machine learning techniques (e.g. SVM and decision trees) exploiting low-level hand-crafted features, such as color, texture, or morphology [6], [18], [22], [25], [29], [30].

Most methods in literature for the analysis of histological images are based on DL networks (e.g., AlexNet, ResNet and UNet [11], [15], [21]), that automatically learn features that optimally represent the data for the problem at hand. In some cases DL networks are integrated with well-known classifiers (e.g., SVM, Random Forest and Adaboost), and the methods adopt appropriate training strategies (e.g., defined patches, pre-processing, parameter setting and selected loss function). Additionally, for a specific task, the performance of the different approaches depends on the adopted network and the developed training strategy. It is no coincidence that many teams in international competitions have adopted the same network architecture on the same dataset with different training strategies for a specific task, obtaining widely different results [3], [17], [24].

Most of these methods are based on the use of Convolutional Neural Networks (CNN) for the detection and classification tasks. In the context of the detection task, the method in [26] won the first place at the ISBI2016 Metastasis Detection Challenge [17], by adopting a CNN with 27 layers, while the method in [8] outperformed in terms of accuracy the other methods at the ICPR2014 MITOS-ATYPIA Challenge [1] by using deep cascaded CNN. For the detection of IDC, the method in [9] yielded the best quantitative results in comparison with approaches using hand-crafted image features on the dataset available for download at [2]. For the classification task, a CNN based on AlexNet was adopted in [23] in order to discriminate between benign and malignant breast cancer tumors. This approach outperformed the previously reported results obtained by other machine learning models trained with hand-crafted textural descriptors on the BreaKHis dataset [22]. The authors in [4] proposed a comparison between two approaches (hand-crafted and CNN based) for the classification of breast cancer histological images, showing that their CNN architecture outperforms the state of the art on BreakHis dataset. Other interesting experimental results were obtained for the same dataset by the method proposed in [5] by classifying breast cancer histological images independently of the image magnification factor. Three different configurations of ResNet were used by the authors in [7] for the multi-classification of breast cancer obtaining a remarkable performance on the images provided for the ICIAR2018 BACH Cahllenge [3]. Finally, in [13] AlexNet, using the same training strategy, was adopted both for IDC detection and lymphoma multi-classification on the datasets available for download at [2], outperforming the method proposed in [9], in relation to IDC detection.

Differently from the methods based on the CNN trained end-to-end technique for classification, other approaches have been based on unsupervised networks, mainly autoencoders (AE), that do not required labeled samples to detect the inner structure to be used for the subsequent detection and classification tasks. Indeed, AEs are commonly used for the pre-training of different deep neural networks or classifiers. An AE is optimized to learn the principal components of the data distribution. However, when a non-linear activation function is used, AE learns over complete unsupervised representations by reconstructing the original input, operating under several constraints (sparsity or hierarchicality) [19]. The authors in [10] employed sparse AEs to learn an unsupervised representation that feeds a softmax classifier over this representation identifying the image regions that are most relevant for the basal cell carcinoma cancer detection. In [12], a Convolutional Sparse AE for simultaneous nucleus detection and feature extraction in histological tissue images was proposed. The foreground image was reconstructed by certain vectors in feature maps that represent salient objects. Additionally, a Stacked Sparse AE framework was presented in [28] for automated nucleus detection on breast cancer histological images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access



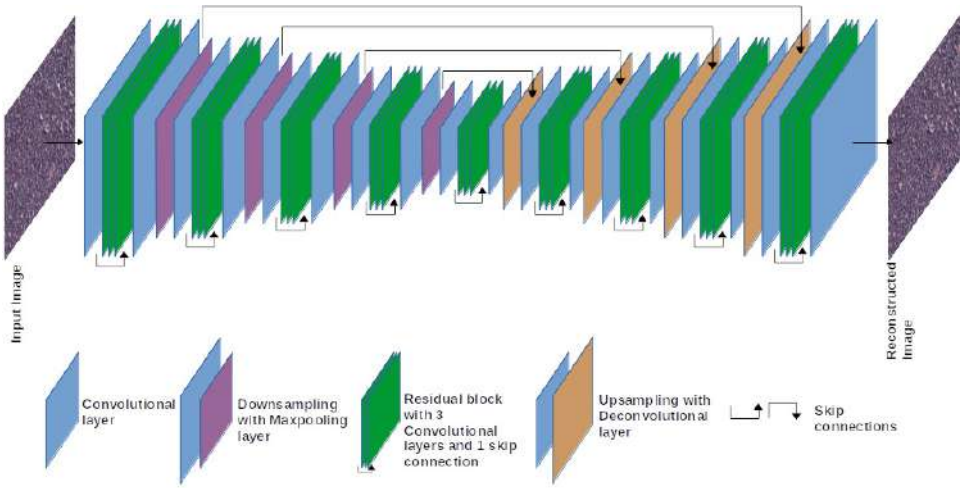



Fig. 1: FusionNet Architecture

## III. Methodology

The proposed approach is designed for IDC detection of and lymphoma multi-classification in H&E histological images. Differently from the segmentation process, the detection task does not involve the delineation of accurate boundaries for the regions of interest, but only the identification of the areas including such regions. For this reason, the detection of IDC can be addressed as a classification problem: the WSI is divided into patches and the final detection is obtained by marking patches with an IDC or a non-IDC label.
The classification for both tasks is based on the use of a convolutional autoencoder (CAE), namely FusionNet [20]. Similarly to all CAE networks, the architecture of FusionNet has a configuration which is completely symmetrical, but it is also a residual network, due to the presence of skip connections (see Fig. 1).

FusionNet introduces long skip connections between the feature maps in the encoder and those located at the same level in the decoder; moreover, short skip connections are present in each residual block of the network. With such a configuration, the information flows within and accross different levels of the network.

We propose two different scenarios for the classification:

1) Classification by reconstruction - the CAE is trained under a sparsity constraint in an unsupervised manner;
2) Supervised Classification - only the encoding part of the CAE is trained and this in a supervised manner.

In both cases, a softmax classifier takes the extracted features as the input. The output of the softmax classifier produces a value between 0 and 1 that can be interpreted as the probability of the input belonging to a given class. The classifier is trained minimizing the Cross Entropy Loss (CE), used to measure the divergence between the effective class $c$ and the predictive class $\hat{c}$ of $n$ samples:

$$CE(c, \hat{c}) = -\sum_{i=1}^{n} c_i \cdot log(\hat{c}_i) \tag{1}$$

In the first scenario, the network is trained in an unsupervised manner extracting features useful for the reconstruction of the input image. The obtained representation of the encoder has a lower dimensionality than the input data. The set of weights associated with the representation can be interpreted as the set of feature maps learned by the CAE to be used for the classification. The Mean Squared Error Loss (MSE) is used to measure the error between the input image $x$ and the reconstructed image $\hat{x}$:

$$MSE(x, \hat{x}) = \frac{\sum_{i=1}^{n}(x_i - \hat{x})^2}{n} \tag{2}$$

Backpropagation is implemented by using the Stochastic Gradient Descent (SGD) algorithm [16], with a controlled learning rate. In general, to improve the performance of an AE and to prevent overfitting, the addition of a sparsity constraint during the training is suitable [19], [28]. The sparsity constraint is imposed on the hidden units, enabling the AE to discover interesting structures in the data. We have imposed the sparsity constraint through the introduction of Kullback-Leibler divergence (KL), that measures the degree of difference between two distributions with means $\rho$ and $\hat{\rho_j}$. In detail, $\rho$ refers to the target activation function of the hidden units and $\hat{\rho_j}$ refers to the average activation function of the hidden unit $j$. The activation fuction used for the hidden units is the sigmoid function.
After the training, a maxpooling layer is applied to the extracted features and the output is passed to a fully connected layer that performs the classification by means of a softmax activation function. In this case, backpropagation is implemented by using the Adaptive Moment Estimation (Adam) algorithm [14], an extension of the SGD. The algorithm computes adaptive learning rates for each network parameter from estimates of first and second moments of the gradient.

In the second scenario, only the encoding part of the CAE is trained in a supervised manner, i.e. the input to the network is represented by the image and the corresponding class. The network is trained end-to-end to learn filters and to combine features with the aim of feeding a fully connected

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access



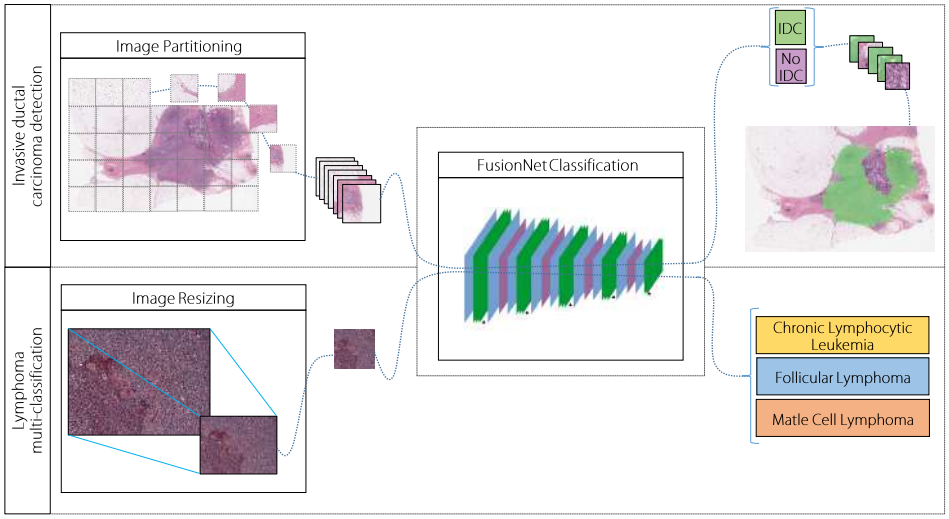



Fig. 2: A graphical representation of SEF

layer. Also in this scenario, the Adam algorithm is used for backpropagation and a softmax activation function is used for the classification.

## IV. Use Cases

The proposed approach is based on FusionNet. In particular, in the "Supervised Classification" scenario only the encoding part of FusionNet is trained, and therefore we refer to this approach as Supervised Encoder FusionNet (SEF). In Fig. 2, a graphical representation of SEF is shown. Moreover, both UNet and ResNet have been considered for comparison in the experiments. In the "Supervised Classification" scenario, we will refer to use of UNet as Supervised Encoder UNet (SEU), while no distinction will be needed for ResNet as it is not considered in the "Classification by reconstruction" scenario.

The performances have been evaluated on two datasets available for download at [2]. In the following section, we will refer to the datasets for IDC detection and for lymphoma multi-classification as D-IDC and D-Lymph, respectively.
Regarding the implementation details, the framework has been implemented in Pytorch on a workstation equipped with 2 Xeon 10-Core E5-2630v4 2,2Ghz 25MB and 4 NVIDIA GEFORCE GTX 1080Ti 11GB PCI-EX.
In the following section, for each use case, details about the adopted dataset, the training strategy, and the experimental results compared with other approaches will be given.

### A. Invasive Ductal Carcinoma Detection

The experiments have been performed on the D-IDC dataset, which includes 162 WSI acquireFFTd at $40\times$, each partitioned into a set of patches with a size equal to $50 \times 50$ pixels. The number of patches representing IDC and non-IDC is $46,633$ and $124,011$, respectively. An example of invasive ductal carcinoma is shown in Fig. 3.

As shown in Fig. 1, the FusionNet encoder is composed of a set of blocks including downsampling layers. Whenever downsampling is performed, the input size is halved. This implies that the input image should have a size allowing for a scaling of factor 2 for each downsampling and guaranteeing that the input size of the bridge layer is not too small. In order to allow a scaling of factor 2, the input of the network is represented by the central square region of a size of $48 \times 48$ pixels, extracted from each patch. Since the encoder is constituted by 4 downsampling layers, the input size of the bridge layer is equal to $3 \times 3$.

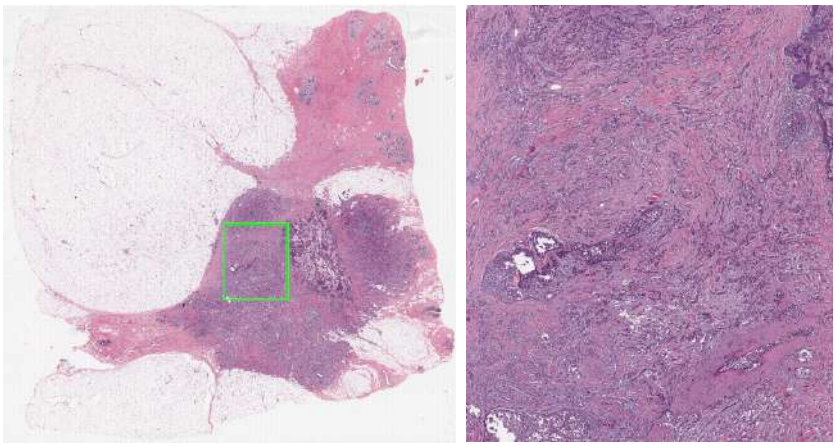
Since a large number of images are available, no augmentation operation has been performed. In order to allow for a comparison with the state of the art, the same training and testing sets as [9], [13] have been used. In particular, the training set consists of about $70\%$ of the whole dataset. The final detection on each WSI will be given in terms of IDC or non-IDC patches.

The values of the parameters adopted for the training are given in Table I. Both scenarios produce 512 feature maps, subsequently used for the classification step. In the "Classification by reconstruction" scenario, a max pooling layer with a kernel size equal to $2 \times 2$ and a stride equal to 2 is applied at the end of the encoding stage, while in the second scenario the max pooling layer is substituted by a fully connected layer.

*1) Experimental Results and Discussion:* We have compared the performance of the different approaches in terms of standard metrics, namely Accuracy, F-Measure, Precision, Sensitivity and Specificity. We evaluated the performance also in terms of the Balanced Accuracy (BAC) measure, calculated as the average between the Specificity and Sensitivity. The numerical results of these experiments are reported in Table II. Depending on the scenario, the performance has been evaluated by using the same values as the training parameters for all the considered approaches with the exception of the number of epochs for the fine-tuned ResNet that was been reduced to 10 to prevent the network from overfitting. We have performed many experiments with different configurations of ResNet (i.e. ResNet-18, ResNet-34 and ResNet-50, fine-tuned and from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access





(a) (b)

Fig. 3: Invasive ductal carcinoma: (a) whole slide image; (b) a magnification of the highlighted box in (a)

| Scenario | $N^{\circ}$ Epochs | Mini-batch | Back. alg. | Init. Lear.rate | Lear.rate update |
|---|---|---|---|---|---|
| Classification by reconstruction | 20 | 64 | SGD | 0.001 | 0.0001 after $15^{th}$ epoch |
| Supervised Classification | 15 | 64 | ADAM | 0.0001 | update by ADAM |

TABLE I: The training parameters for the IDC detection

| Scenario | Network | Acc. | F-meas. | Prec. | Sens. | Spec. | BAC |
|---|---|---|---|---|---|---|---|
| Classification by reconstruction | FusionNet | 84.71 | 71.70 | 73.07 | 70.38 | 90.15 | 80.26 |
| | UNet | 84.65 | 68.30 | 79.14 | 60.07 | **93.99** | 77.03 |
| Supervised Classification | SEF | **89.57** | **81.54** | **79.45** | **83.75** | 91.77 | **87.76** |
| | SEU | 85.87 | 71.90 | 79.44 | 65.67 | 93.54 | 79.60 |
| | ResNet-34 | 87.24 | 77.14 | 76.08 | 77.22 | 90.66 | 84.44 |

TABLE II: Results of the different approaches for the IDC detection. The best performance is highlighted in bold

| Network | Acc. | F-meas. | Prec. | Sens. | Spec. | BAC |
|---|---|---|---|---|---|---|
| ResNet-18 fine-tuned | 86.79 | 75.38 | 77.39 | 73.48 | 91.85 | 82.67 |
| ResNet-18 from scratch | 85.28 | 72.94 | 73.83 | 72.07 | 90.30 | 81.18 |
| ResNet-34 fine-tuned | 87.24 | **77.14** | 76.08 | **77.22** | 90.66 | **84.44** |
| ResNet-34 from scratch | 85.11 | 72.32 | 73.65 | 70.69 | 90.42 | 80.55 |
| ResNet-50 fine-tuned | **87.52** | 77.02 | **78.03** | 76.08 | **91.87** | 83.97 |
| ResNet-50 from scratch | 85.00 | 72.91 | 73.37 | 72.47 | 90.01 | 81.24 |

TABLE III: Results of the different ResNet configurations for the IDC detection. The best performance is highlighted in bold

| Scenario | Network | TP | TN | FP | FN |
|---|---|---|---|---|---|
| Classification by reconstruction | FusionNet | 9964 | 33610 | 3672 | 4193 |
| | UNet | 8504 | **35041** | **2241** | 5653 |
| Supervised Classification | SEF | **11857** | 34215 | 3067 | **2300** |
| | SEU | 9297 | 34876 | 2406 | 4860 |
| | ResNet-34 | 11074 | 33800 | 3482 | 3083 |

TABLE IV: TP, TN, FP and FN calculated by the different approaches for the IDC detection. The best performance is highlighted in bold

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access



| | IDC | | Lymphoma |
|---|---|---|---|
| | F-meas. | BAC | Acc. |
| SEF | **81.54** | **87.76** | **97.67** |
| Method in [13] | 76.48 | 84.68 | 96.58 |
| Method in [9] | 71.80 | 84.23 | - |

TABLE V: The comparisons with the state of the art for the IDC detection and lymphoma multi-classification. The best performance is highlighted in bold

scratch). The results are shown in Table III. Independently of the configuration, ResNet always outperformed FusionNet, UNet and SEU, following SEF in the ranking. For the sake of simplicity, we report in Tables II only the results for the best configuration of ResNet in terms of the F-measure and BAC (i.e. the fine-tuned ResNet-34).

The SEF method significantly outperformed the other DL approaches in terms of the accuracy. Indeed, it achieved an overall increment of $4.86\%$, $5.52\%$, $3.7\%$ and $2.33\%$ in accuracy as compared to FusionNet, UNet, SEU and ResNet-34, respectively. Also for the remaining measures, SEF proved to be the best configuration, followed in the ranking by ResNet-34, with the exception of the performance in terms of Specificity, where UNet turns out to be the winning approach.

The numbers of true positive patches (TP), true negative patches (TN), false positive patches (FP) and false negative patches (FN) are reported in Table IV.

Figs. 4 and 5 illustrate examples of the IDC detection result of SEF versus the FusionNet, UNet, SEU and ResNet methods compared to the ground truth. The true positive patches are highlighted in green, while the false positive patches are highlighted in red. All methods showed a reasonable detection performance, but SEF (Figs. 4(d) and 5(d)) revealed much more information compared to the other images, the result being closer to the ground truth. False positive patches produced by FusionNet and ResNet-34 tend to accumulate in specific regions, while those generated by SEF are sparsely distributed. This suggests that SEF allows to adopt such a kind of region based decision rules (e.g. the majority voting), such that a patch is definitely classified as IDC only when this classification label extends over multiple adjacent patches.

Finally, in the first column of the Table V comparisons between SEF and the other two methods ( [13] and [9]) in the literature are reported. Methods [13] and [9] were evaluated using the same dataset D-IDC and are based on AlexNet and a 3-layered CNN, respectively. They were ranked according to according to the performance of the F-measure and BAC. The SEF approach outperformed method [13] with an increment of $5.06\%$ and $3.08\%$ in terms of the F-measure and BAC, respectively. With respect to method [9], the performance of SEF showed an increment of $9.74\%$ and $3.53\%$ in terms of the F-measure and BAC measure, respectively. Comparing the results of Tables II and V, it is clear that ResNet-34 also outperformed both methods [13] and [9] in terms of the F-measure, while the BAC measure was approximately equivalent in all these methods. Finally, residual CNNs end-to-end trained for classification (i.e., SEF and ResNet) provided the best performances for the IDC detection task. In particular, SEF provided the best performance.

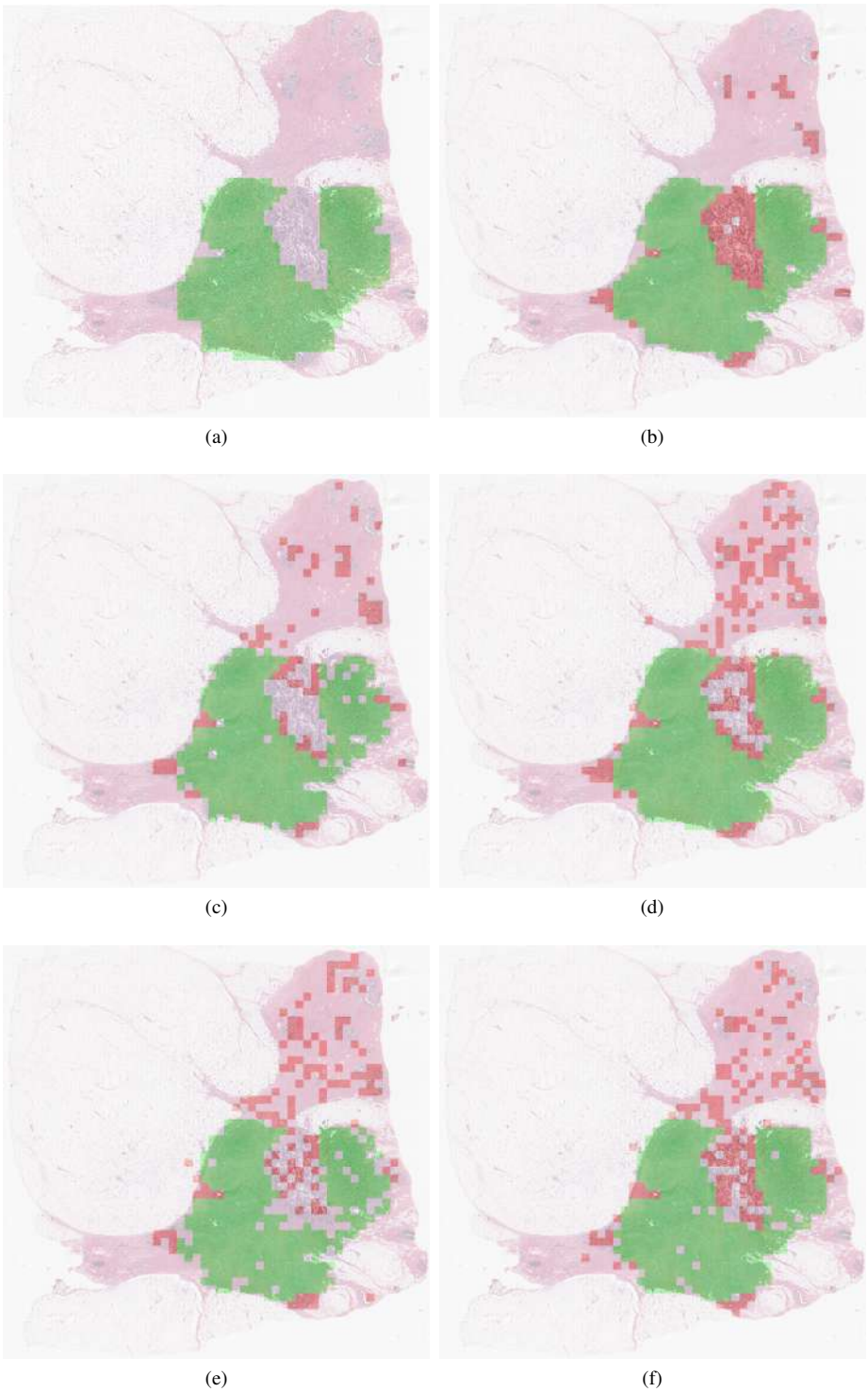
### B. Lymphoma Multi-classification

The D-Lymph dataset has been employed as a benchmark to evaluate the image analysis techniques for the CLL, FL and MCL sub-types. In total $374$ images were generated, containing $113$, $139$, and $122$ images of CLL, FL and MCL, respectively. The size of each image was equal to $1388 \times 1040$ pixels. In Fig. 6 a sample for each class is shown. With high resolution images, DL models suffer from high computational costs and limitations in the number of model layers and channels. In accordance with a general trend, the authors in [13] address this problem by training their DL network on image patches and classifying an image based on patch-level predictions. We decide on a different strategy, that of reducing the input image by a factor about $88\%$. Doing so, the spatial organization of cellular structures can be globally analyzed by the network, while also keeping the computational cost down. The reduced image had a size equal to $170 \times 128$ pixels. We considered the central square section with a size equal to $128 \times 128$ pixels as the input to the network. In this case, the input size of the bridge layer was equal to $8 \times 8$. An augmentation of the dataset was performed by considering a set of 9 transformations for each training image (i.e. horizontal flip, vertical flip, three clock-wise rotations of $90°$, and two horizontal and two vertical translations of $\pm 100$ pixels). In order to allow a comparison with [13], the same experimental protocol was followed. In particular, in [13] a k-fold cross validation with $k = 5$ was adopted, where each folder contained $299$ training images (consisting of about $80\%$ of the whole dataset) and $75$ test images. Differently from [13], each folder contained resized central regions of the image together with their corresponding augmentations, instead of patches resulting from a splitting operation. The values of the parameters adopted for the training are given in Table VI. Both scenarios produced $512$ feature maps, at the end of the encoding stage. In the "Classification by Reconstruction" scenario, a max pooling layer with a kernel size equal to $4 \times 4$ and a stride equal to $2$ was applied at the end of the encoding part, producing $2048$ feature maps. In the "Supervised Classification" scenario, the max pooling layer was subsituted by a fully connected layer, which did not produce any increase in the number of feature maps.

*1) Experimental Results and Discussion:* The performances of the different approaches have been quantitatively assessed by considering the Accuracy measure. The numerical results of these experiments are reported in Table VII.
Also in this case, we have performed many experiments with different configurations of ResNet (i.e. ResNet-18, ResNet-34 and ResNet-50, fine-tuned and from scratch and by using two different input sizes i.e. $128 \times 128$ and $224 \times 224$ pixels). Moreover, by performing Wilcoxon test [27] to validate the statistical significance, we found that the obtained p-values always exceed $0.7$. For this reason, we preferred to report in Table VIII the results produced by the fine-tuned ResNet-34, as it provides the lowest standard deviation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access



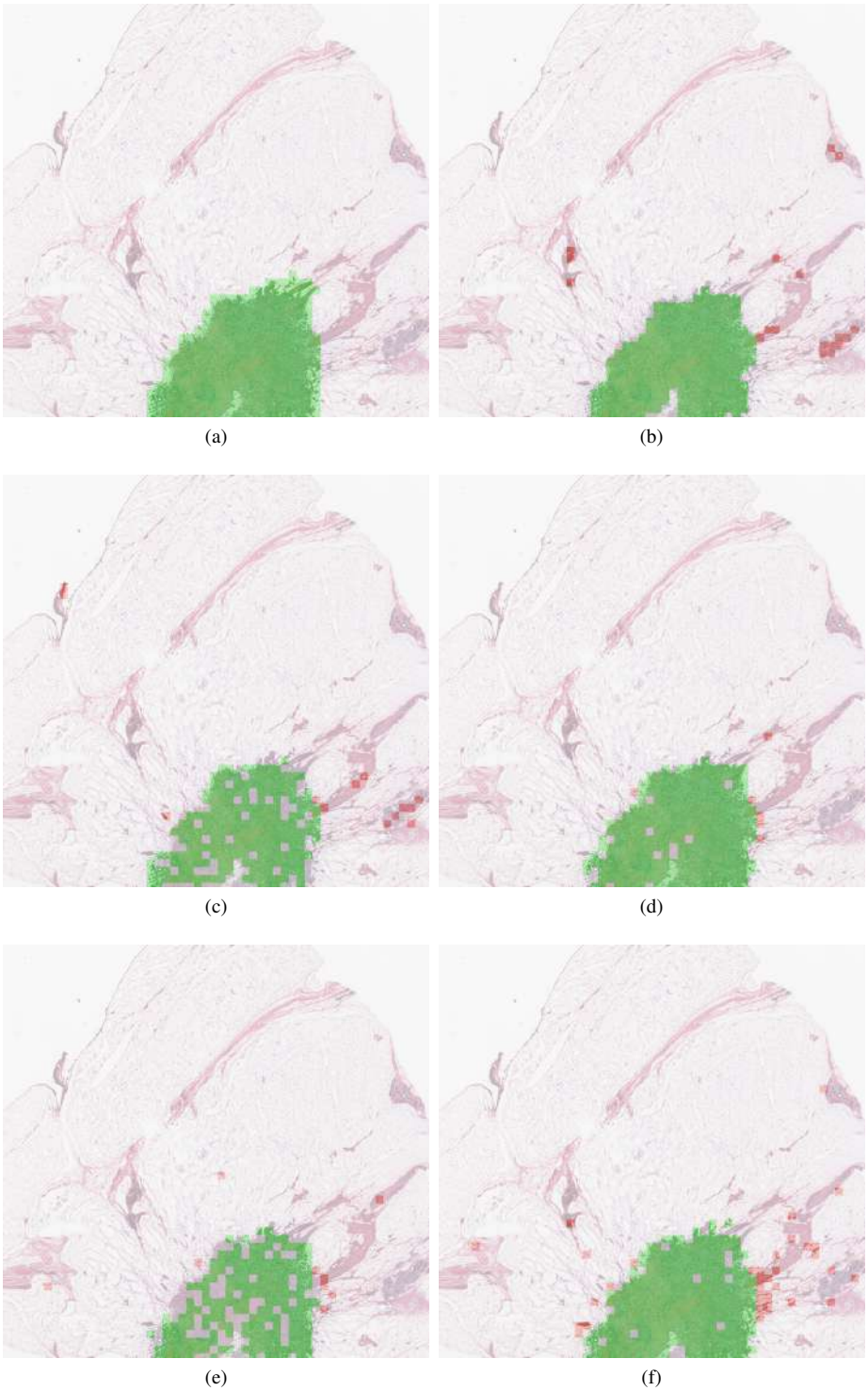


(a) (b) (c) (d) (e) (f)

Fig. 4: Results for the IDC detection (true positive and false positive patches are highlighted in green and in red, respectively): (a) ground truth (b) FusionNet (c) UNet (d) SEF (e) SEU (f) ResNet-34

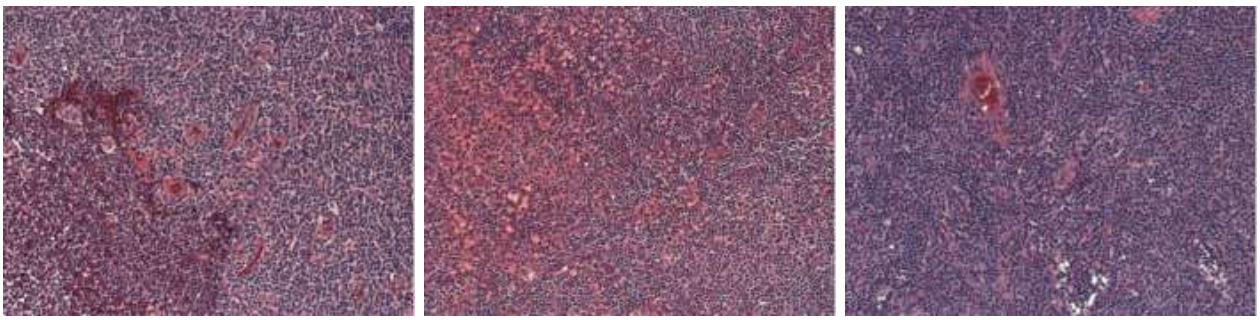
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2908724, IEEE Access





(a) (b) (c) (d) (e) (f)

Fig. 5: Results for the IDC detection (true positive and false positive patches are highlighted in green and in red, respectively): (a) ground truth (b) FusionNet (c) UNet (d) SEF (e) SEU (f) ResNet-34



(a) (b) (c)

Fig. 6: Lymphoma sub-types: (a) MCL (b) CLL (c) FL

| Scenario | $N^\circ$ Epochs | Mini-batch | Back. alg. | Init. Lear.rate | Lear.rate update |
|---|---|---|---|---|---|
| Classification by reconstruction | 200 | 32 | SGD | 0.001 | 0.0001 after $150^{th}$ epoch |
| Supervised Classification | 150 | 32 | ADAM | 0.0001 | update by ADAM |

TABLE VI: The training parameters for the lymphoma multi-classification

| Scenario | Network | Acc. | St. Dev. |
|---|---|---|---|
| Classification by reconstruction | FusionNet | 77.60 | 5.4 |
| | UNet | 65.10 | 8.4 |
| Supervised Classification | SEF | **97.67** | **1.3** |
| | SEU | 92.80 | 3.2 |
| | ResNet-34 | 95.47 | 2.0 |

TABLE VII: Results of the different approaches for the lymphoma multi-classification. The best performance is highlighted in bold

| Network | Acc. |
|---|---|
| ResNet-18 fine-tuned ($128 \times 128$) | 93.60 |
| ResNet-18 from scratch ($128 \times 128$) | 81.87 |
| ResNet-18 fine-tuned ($224 \times 224$) | **95.73** |
| ResNet-18 from scratch ($224 \times 224$) | 83.20 |
| ResNet-34 fine-tuned ($128 \times 128$) | 93.86 |
| ResNet-34 from scratch ($128 \times 128$) | 80.00 |
| ResNet-34 fine-tuned ($224 \times 224$) | 95.47 |
| ResNet-34 from scratch ($224 \times 224$) | 81.86 |
| ResNet-50 fine-tuned ($128 \times 128$) | 93.07 |
| ResNet-50 from scratch ($128 \times 128$) | 86.93 |
| ResNet-50 fine-tuned ($224 \times 224$) | 94.93 |
| ResNet-50 from scratch ($224 \times 224$) | 87.47 |

TABLE VIII: Results of the different ResNet configurations for the lymphoma multi-classification. The best performance is highlighted in bold

| Method | p-value |
|---|---|
| FusionNet | 0.0088 |
| UNet | 0.0088 |
| SEU | 0.0142 |
| ResNet-34 | 0.1288 |

TABLE IX: The p-values provided by the Wilcoxon test between SEF and each other methods

| | MCL | CLL | FL |
|---|---|---|---|
| MCL | 16.8 | 4.4 | 3.2 |
| CLL | 5.2 | 15.6 | 1.8 |
| FL | 0.8 | 1.4 | 25.6 |

TABLE X: The confusion matrix of FusionNet

| | MCL | CLL | FL |
|---|---|---|---|
| MCL | 13.4 | 2.8 | 8.4 |
| CLL | 8 | 10.8 | 3.8 |
| FL | 2.6 | 0.6 | 24.6 |

TABLE XI: The confusion matrix of UNet

The SEF method achieved an accurancy of $97.67\%$ which is $20.07\%$, $32.57\%$, $4.87\%$, and $2.2\%$ higher when compared with those of FusionNet, U-Net, SEU and ResNet-34, respectively. In this use case also, the SEF and ResNet networks outperformed the other networks. The same values for the training parameters were used for all the considered approaches, with the exception of fine-tuned ResNet-34, which adopted a lower

| | MCL | CLL | FL |
|---|---|---|---|
| MCL | 23.6 | 0.8 | 0.2 |
| CLL | 0.4 | 22 | 0.2 |
| FL | 0.2 | 0 | 27.6 |

TABLE XII: The confusion matrix of SEF

| | MCL | CLL | FL |
|---|---|---|---|
| MCL | 20.4 | 3.2 | 1 |
| CLL | 1.8 | 20.2 | 0.6 |
| FL | 0.8 | 0.4 | 26.6 |

TABLE XIII: The confusion matrix of SEU

| | MCL | CLL | FL |
|---|---|---|---|
| MCL | 22.4 | 1 | 1.2 |
| CLL | 0.6 | 21.8 | 0.2 |
| FL | 0.4 | 0 | 27.4 |

TABLE XIV: The confusion matrix of ResNet-34

number of epochs (50 epochs) and a different resolution of the input images ($224 \times 224$ pixels). In order to further strengthen these results, we have also performed Wilcoxon test between SEF and the other methods. The resulting p-values are reported in table IX. With the exception of the comparison between SEF and Resnet-34 (p-value=0.1288), the obtained p-values are always lower than 0.01. Other observations can be made by looking in Table VII at the values of the standard deviation of accuracy, computed by adopting a 5-fold cross validation. The SEF method had a much lower standard deviation of accuracy when compared with those of FusionNet, UNet, SEU and ResNet-34. Thus, the performance of SEF had a lower dependence on the selection of the training set than each of the other approaches. Additionally, the Tables X, XI, XII, XIII and XIV show the confusion matrices of Fusionet, UNet, SEF, SEU and ResNet-34, respectively. Considering that the average number of MCL, CLL and FL images, in the test set, is equal to 24.6, 22.2 and 27.8, respectively, by means of an analysis of the confusion matrices it is clear that the number of miss-classified images with SEF was considerably lower than those provided by the other approaches. The performance of SEF is compared with method [13], which adopts AlexNet also for this use case and it is evaluated only in terms of accuracy (see the last column of the Table V). The SEF method achieved an accurancy of 97.67% which was 1.09%, higher when compared with the accuracy of method [13].

For the sake of completeness, we tested also the same training strategy as [13] for SEF and ResNet-34. Instead of image resizing preprocessing, the images were split into $48 \times 48$ patches with a stride of 48 avoiding the augmentation of the dataset. During the testing stage, patches were extracted using the same methodology, and a voting scheme per subtype was used where the votes were aggregated to predict the class. In particular, the class with the highest number of votes became the detected class for the entire image. According to this training strategy, ResNet-34 achieved an accuracy equal to 96.84% outperforming method [13]. SEF obtained an accuracy equal to 97.06% and it proved once again to have the best performance.

## V. Conclusions

In this work, we have suggested a method, namely SEF, based on a deep network for learning histological images to avoid hand-crafted pathological features. Using deep learning approaches with specific settings for cancer detection and classification is an effective and reliable strategy compared to conventional approaches. We have shown that the encoder of FusionNet, which has been designed for image segmentation and reconstruction, can be adapted for cancer detection and the classification of histological images.

In detail, our SEF method is based on a Residual CNN (i.e. the encoder of FusionNet) and a softmax classifier to address two use cases: the detection of IDC of the breast cancer and lymphoma multi-classification.

In our experiments, we compared the performances of SEF against different existing deep neural network (FusionNet, UNet and ResNet) and the encoding part of UNet under the same conditions and on the same datasets.

We have also proposed several strategies for training of the proposed network, based on the extraction of patches or on resized images, allowing us to deal with the high-resolution of histological images. The detailed experimental analysis and performance comparisons show a significant improvement of the SEF method in relation to all the considered DL approaches and other existing methods for both use cases. The results show that for the considered uses cases, Autoencoders (FusionNet and UNet) extract features that are unsuitable for the classification, as they are learned for the image reconstruction. This is the underlying reason why CNNs trained end-to-end for classification have provided a higher performance. In particular, the residual one (SEF and RESNet) are better at consider for small cellular structures depicted in the histological images, as they attenuate the drawback of vanishing gradients. However, this problem remains, so increasing the deep of the network (from 34 to 50 layers in RESNet) does not translate into an increasing of the classification performance. In our future work we aim to explore the application of SEF to other use cases for histological image analysis and also with different training strategies.

## References

[1] "2014 ICPR MITOS-ATYPIA." [Online]. Available: http://mitosatypia-14.grand-challenge.org/

[2] "Lymphoma and IDC datasets." [Online]. Available: http://www.andrewjanowczyk.com/deep-learning/

[3] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan *et al.*, "Bach: Grand challenge on breast cancer histology images," 2018.

[4] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24 680–24 693, 2018.

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

[6] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2016.

[7] N. Brancati, M. Frucci, and D. Riccio, "Multi-classification of breast cancer histology images by using a fine-tuning strategy," in *Image Analysis and Recognition*, A. Campilho, F. Karray, and B. ter Haar Romeny, Eds. Cham: Springer International Publishing, 2018, pp. 771–778.

[8] H. Chen, Q. Dou, X. Wang, J. Qin, P.-A. Heng *et al.*, "Mitosis detection in breast cancer histology images via deep cascaded networks." in *AAAI*, 2016, pp. 1160–1166.

[9] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*, vol. 9041. International Society for Optics and Photonics, 2014, p. 904103.

[10] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 403–410.

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

[12] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran *et al.*, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognition*, vol. 86, pp. 188–200, 2019.

[13] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, vol. 7, 2016.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

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

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, "Detecting cancer metastases on gigapixel pathology images," 2017.

[18] R. Nateghi, H. Danyali, and M. S. Helfroush, "Maximized inter-class weighted mean for fast and accurate mitosis cells detection in breast cancer histopathology images," *Journal of medical systems*, vol. 41, no. 9, p. 146, 2017.

[19] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, 2011.

[20] T. M. Quan, D. G. Hilderbrand, and W.-K. Jeong, "Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics," 2016.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[23] ——, "Breast cancer histopathological image classification using convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 2560–2567.

[24] M. Veta, P. J. Van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. Larsen, J. S. Vestergaard, A. B. Dahl *et al.*, "Assessment of algorithms for mitosis detection in breast cancer histopathology images," *Medical image analysis*, vol. 20, no. 1, pp. 237–248, 2015.

[25] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE transactions on medical imaging*, vol. 35, no. 3, pp. 738–751, 2016.

[26] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[27] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[28] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[29] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, p. 12474, 2016.

[30] Y. Zhang, B. Zhang, F. Coenen, J. Xiao, and W. Lu, "One-class kernel subspace ensemble for medical image classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 17, 2014.