International Journal of Cognitive Informatics and Natural Intelligence
Volume 14 • Issue 4 • October-December 2020

Efficient Regularization Framework for Histopathological Image Classification Using Convolutional Neural Networks.

Nassima Dif, EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
https://orcid.org/0000-0002-8683-3163

Zakaria Elberrichi, EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
https://orcid.org/0000-0002-3391-6280

ABSTRACT

Deep learning methods are characterized by their capacity to learn data representation compared to the traditional machine learning algorithms. However, these methods are prone to overfitting on small volumes of data. The objective of this research is to overcome this limitation by improving the generalization in the proposed deep learning framework based on various techniques: data augmentation, small models, optimizer selection, and ensemble learning. For ensembling, the authors used selected models from different checkpoints and both voting and unweighted average methods for combination. The experimental study on the lymphomas histopathological dataset highlights the efficiency of the MobileNet2 network combined with the stochastic gradient descent (SGD) optimizer in terms of generalization. The best results have been achieved by the combination of the best three checkpoint models (98.67% of accuracy). These findings provide important insights into the efficiency of the checkpoint ensemble learning method for histopathological image classification.

KEYWORDS

Convolutional Neural Network, Deep Learning, Ensemble Learning, Histopathology, Lymphoma, MobileNet

1. INTRODUCTION

The histopathology is a branch of histology where biological diseased tissues or cells are examined under a microscope. Usually, the pathologist observes the stained biopsies with hematoxylin and eosin (H&E) for prognosis, grading, and cancer identification. Nevertheless, the diagnostic of biopsies is a complex task and requires years of experience, which results a high variance between the pathologist's diagnosis. Thus, to reduce this inter variability, computer-aided diagnostic systems (CAD) are employed as a second reader.

Previously, machine learning (ML) methods have been one of the most popular applications in CAD systems, their overall process depends on four main steps: the detection of regions of interest, feature extraction (GLCM (Vujasinovic et al., 2015), LBP (Hervé et al., 2011)), feature selection (RFA (Dif et al., 2019), MVO (Dif et al., 2018)) and classification (Komura et al., 2018). However, in histopathology, the extraction of handcrafted features is one of the greatest challenges because of the complex structure of cells and tissues, moreover, tumors can present distinct cytological features (Roberto et al., 2017). Recently, there has been a surge of interest in deep learning (DL) algorithms

DOI: 10.4018/IJCINI.2020100104

for medical image analysis. The benefit of these methods compared to the traditional ML algorithms is their capacity to learn data representation, where the relevant characteristics are extracted through the classification process.

For histopathological image analysis, biopsy slides are digitized as whole slide images (WSI) (Janowczyk et al., 2016) by whole slide digital scanners (WSD) (Al-Janabi et al., 2012). The high resolution of WSI made digital pathology a popular application for DL methods (Litjens et al., 2017) in different tasks: mitosis detection (Cireşan et al., 2013), gland segmentation (Kainz et al., 2015), lymphoma subtype classification (Janowczyk et al., 2016). On the other hand, the limited number of available medical images and the difficulty of their annotation are the leading causes of overfitting. In the literature, several attempts have been made to prevent this problem by improving the generalization based on various techniques: transfer learning (Ng et al., 2015) and regularization strategies (dropout (Srivastava et al., 2014) and ensemble learning (Ju et al., 2018)).

The purpose of this research is to improve the generalization capacity of convolutional neural networks for lymphoma subtypes classification. Our work takes advantage of various regularization methods in a deep learning framework: data augmentation, the exploitation of small models (MobileNet), the selection of the suitable optimizer, and the checkpoint ensemble model selection. This study provides the first comprehensive assessment of checkpoint ensembling in histopathological applications based on the optimized MobileNet architecture.

The remaining part of the paper proceeds as follows: section1 presents the related works to the automated methods for lymphoma subtypes classification and the ensemble learning methods in deep learning. Section 2 details the process of the used methods. Section3 explains the proposed framework. Section 4 illustrates and discusses the obtained results and the last part concludes this work.

2. RELATED WORKS

2.1. Lymphomas Subtypes Classification

Lymphomas are tumors affecting lymphocytes (T-or B-cells) (Orlov et al., 2010). They are classified into Non-Hodgkin's lymphomas (NHL) and Hodgkin's lymphomas (HL) (Chan et al., 2001). The NHL represents 90% of lymphomas (Shankland et al., 2012). It includes different subtypes such as Diffuse large B-cell lymphoma (DLBCL), which is the most common form, the chronic lymphocytic leukemia (CLL), follicular lymphoma (FL) and mantle cell lymphoma (MCL). These subtypes present aggressive (DLBC, MCL) or indolent (CLL, FL) NHLs, where the aggressive form progress rapidly compared to the indolent form.

Several works have been interested in the NHL segmentation (Tosta et al., 2017) and classification. This work aims to enhance the NHL subtypes classification.

(Shamir et al., 2008) proposed the IICBU benchmark suite, which is composed of 9 biological datasets, where the size of images varied from 25 x 25 to 1388 x 1040. Their purpose was to support the computer vision experts for proposing accurate methods for biological images classification. Their experimental study based on the weighted neighbor distance (WND-CHARM) approach has proved the degree of complexity of the automated lymphomas classification, which encouraged the computer vision community to develop more robust methods.

Table 1 provides a summary of the literature relating to the NHL subtypes classification. The proposed strategies are categorized into machine and deep learning methods. Previously, most studies in this field have only focused on the ML applications, where the extraction of the handcrafted features has attracted considerable attention because of the complex morphology of histopathological images.

Various strategies have been proposed for feature extraction. For instance, (Orlov et al., 2010) proposed the exploitation of the transform-based global features for classification. First, they transformed the raw pixels into spectral planes, then, different global features have been computed: texture, polynomial, and other statistical features. Finally, the obtained features were filtered by the

Table 1. The proposed methods in the literature for lymphoma subtypes classification

Method	Article	Classifier	Feature extraction	Feature selection
Machine learning	(Shamir et al., 2008)	WND-CHARM		-
	(Meng et al., 2010)	(C-RSPM) + WMVA	Color and texture features	Chi-square
	(Orlov et al., 2010)	WND	Global features (texture, polynomial and statistical)	FLD, mRmR, F/C
	(Di Ruberto et al., 2015)	Support vector machine (SVM)	Modified GLCM	-
	(Nava et al., 2016)	KFDA	Discrete orthogonal moments (DOMs)	RELIEF
	(Song et al., 2016)	SVM	Visual texture descriptors (IFV, LBP, HOG, GIST, CENTRIST)	SDT
	(Tosta et al., 2018)	SVM	Segmentation with GA	-
Machine learning with CNN models as feature extractors	(Codella et al., 2016)	Non-linear SVMs	Low-level features (color histogram, edge histogram, LBP, transferred CNN ImageNet model)	
	(Song et al., 2017)	SVM	Pre-trained VGG-VD	CFV
	(Song et al., 2017b)	SVM	Local features (SHIFT, pre-trained VGG-VD)	SDR
	(Bai et al., 2019)	Random forest	LTP, MLPQ, CLBP, RIC, FBSIF, AHP, GOLD, HOG, MOR, CLM, LET, GoogleNet, VGGNet, ResNet, Inception, IncResv2	-
	(Nanni et al., 2018)	SVM	LBP, texture representations and statistical features, CNNs learned features.	-
Deep learning	(Janowczyk et al., 2016)	AlexNet (Cifar-10 version)		-

feature selection methods: fisher discriminant analysis (FLD), minimum redundancy maximum relevance (mRmR) and fisher correlation (F/C).

In another investigation, (Meng et al., 2010) presented a new approach based on collateral representative subspace projection modeling (C-RSPM). They started by dividing each image into 25 blocks, then, a set of 505 features (color and texture) have been extracted from these blocks. Finally, fifty features were selected by the Chi-square feature selection method. Another investigation (Di Ruberto et al., 2015) has adapted three greyscale texture models to color texture features for feature extraction from the lymphomas colored images. Another example is the work of (Nava et al., 2016) that proposes a combination of discrete orthogonal moments (DOMs) on the pre-trained images by the color deconvolution method. (Song et al., 2016) suggested that the exploitation of the subcategory discriminant transform (SDT) method on the extracted features is important to minimize the within-class variance and to optimize the between-class difference. The main characteristic of the works described above is their independence on the segmentation methods. The proposed investigation by (Tosta et al., 2018) suggests the use of an unsupervised segmentation method based on genetic algorithms (GA) for CLL and FL neoplastic nuclei segmentation, where they employed four fitness functions (Fisher information, entropies of Renyi, Shannon and Tsallis) to evaluate the GA solutions.

According to Table 1, various feature extraction methods have been used, where it was hard to define the suitable feature extraction method. Therefore, to enhance the set of the extracted features, one of the more practical ways was the exploitation of the deep learning methods as feature extractors. For instance, (Codella et al., 2016) proposed a multi-stage visual learning approach, where variable low-level features have been combined with the extracted feature based on a pre-trained CNN model. In another approach, (Song et al., 2017; Song (b) et al., 2017) proposed a supervised intra-embedding method, where the features have been extracted based on the Convnet-based FV (CFV), as an extension, they proposed to combine the resulting vector with other types of local features. Despite the proposed methods that suggest to combine first the set of extracted features and classify them based on one learning algorithm, other investigations classify separately the extracted features by different methods, and then combine the scores of the resulting models based on ensemble learning methods (Bai et al., 2019; Nanni et al., 2018). For instance, (Bai et al., 2019) combined a random forest classifier trained on extracted textural and statistical features and a softmax classifier trained on the extracted features by a pre-trained inceptionNet. Their study has indicated the potential of the used distance matrix weighting (DMW) method to combines the patch-level results. Another example is the work of (Nanni et al., 2018) that proposes a framework of texture descriptors and CNNs learned features, where they combined the obtained scores based on sum rule.

The use of the deep learning methods for classification has received little attention within the lymphoma's subtypes classification task. To the best of our knowledge, (Janowczyk et al., 2016) made the first attempt to train a convolutional neural network from scratch for lymphomas subtypes classification.

2.2. Ensemble Deep Learning

Overfitting presents the high variance between the train and the test sets performances. To reduce this variance, the exploitation of regularization strategies such as dropout (Srivastava et al., 2014), and ensemble learning methods (Ju et al., 2018) is recommended.

Ensemble learning methods combine the decision of several models to improve both performance and generalization. Several attempts have been made to prove the efficiency of ensemble learning on deep neural networks: stacking (Deng et al., 2014), boosting (Mosca et al., 2016), voting (Xu et al., 2017) and averaging (Chen et al., 2016; Krizhevsky et al., 2012; Zeiler et al., 2014; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016). Where, various strategies have been used to generate the set of models by varying in initialization methods, architectures, optimizers, and the training dataset.

The main drawbacks of DL algorithms is their high computational complexity and the material's requirements illustrated in Table 2, which limits the use of the traditional model generalization methods for training more than one model. A reasonable approach to tackle this issue was the exploitation of the iterative learning process of neural networks to produce a checkpoints ensemble within one training process, also known as self-ensemble. These methods have been employed in various domains: text categorization (Wang et al., 2014), translation systems (Sennrich et al., 2016; Vaswani et al., 2017), abstractive summarization (Kobayashi et al., 2018), malware detection (Sang et al., 2018), images classification (Ju et al., 2018), medical images segmentation (Fok et al., 2018; Jung et al., 2018), facial emotion recognition (Sang & Ha, 2018) and large scale video labeling (Skalic et al., 2017). Table 3 summarizes the used DNN architectures in the proposed checkpoint ensemble methods.

Different strategies have been adopted to select the appropriate checkpoints. For instance, (Chen et al., 2017) have averaged between the best three checkpoint models within the set of deep models, where they proved the efficiency of prediction averaging compared to weight averaging between deep neural networks. In the same way, (Fok et al., 2018) combined between the best 2 to 5 models, and (Sang et al., 2018) have averaged between 25 best models. Other major studies suggest combining between the last checkpoints (Sennrich et al., 2016; Vaswani et al., 2017; Ju et al., 2018). For instance, (Sennrich et al., 2016) proposed to combine the last four models that have been saved every 30 000 mini-batch. Similarly, (Vaswani et al., 2017) have averaged between the last 5 and 20 checkpoints saved at 10-minute intervals. Other investigations suggest combining the generated models from the last epochs (Sang & Ha, 2018; Kobayashi et al., 2018).

Table 2. Material requirements and run time to train convolutional neural networks on the ImageNet dataset

Network	Time	Materiel
AlexNet (Krizhevsky et al., 2012)	Five to six days	Two NVIDIA GTX580 3GB GPUs
ZFNet (Zeiler et al., 2014)	12 days	Single NVIDIA GTX580 GPU
Inception (Szegedy et al., 2015)	One week (estimation)	Few high-end GPUs
VGGNet (Simonyan et al., 2014)	23 weeks depending on the architecture.	Four NVIDIA Titan Black GPUs
Xception (Chollet, 2017)	3 days	60 NVIDIA K80 GPUs

Table 3. The previously combined DNN architectures by the checkpoint ensemble method

Reference	Architecture
(Chen et al., 2017)	Vanilla neural network Convolutional neural network Long Short term memory network
(Sennrich et al., 2016)	Encoder-decoder networks
(Vaswani et al., 2017)	Transformer network based on encoder-decoder network
(Ju et al., 2018)	NIN, VGGNet, ResNet
(Sang & Ha, 2018)	DenseNet Mixture of Neural-Network Experts (MoNN)
(Skalic et al., 2017)	Long Short-Term Memory (LSTM) Gated Recurrent Units (GRU)
(Kobayashi et al., 2018)	LSTM encoder/decoder
(Fok et al., 2018)	ResNet34
(Sang et al., 2018)	RNSALL based on the ResNet model

3. NETWORKS AND OPTIMIZERS

3.1. Networks

MobileNet (Howard et al., 2017) is a convolutional neural network designed especially for mobiles and embedded vision applications. The purpose of this architecture is to design small models through the depthwise separable convolution (DSC) modules and thus to optimize both recognition time and memory requirements. Moreover, MobileNet introduces the concepts of width and resolution multipliers to customize the model according to the device restrictions.

The DSC factorizes the convolution into a depthwise (DC) and a pointwise convolution (PC). First, the DC performs a single 3x3 filter per input channel, then the PC combines the output by 1x1 filters. Equation 1 presents the reduction factor in terms of the number of multiplications performed by DSC compared to a standard convolution. N is the number of filters and D_K is the kernel size.

$$\frac{CSD}{C} = \left(\frac{1}{N} + \frac{1}{D_k^2} \right) \tag{1}$$

MobileNetV2 (Sandler et al., 2018) is an extended version of the MobileNet. The idea is to apply depthwise convolutional filters on larger intermediate tensors by expending the convolutional layers to maintain more information. Then, the original size is restored by a projection (1x1 filters). MobileNetV2 is based on the standard residual connections (He et al., 2016) to prevent the vanishing gradient problem. Figure 1 highlights the difference between a standard convolution, a depthwise separable convolution, and the introduced inverted residuals and linear bottlenecks in the mobileNetV2.

3.2. Optimizers

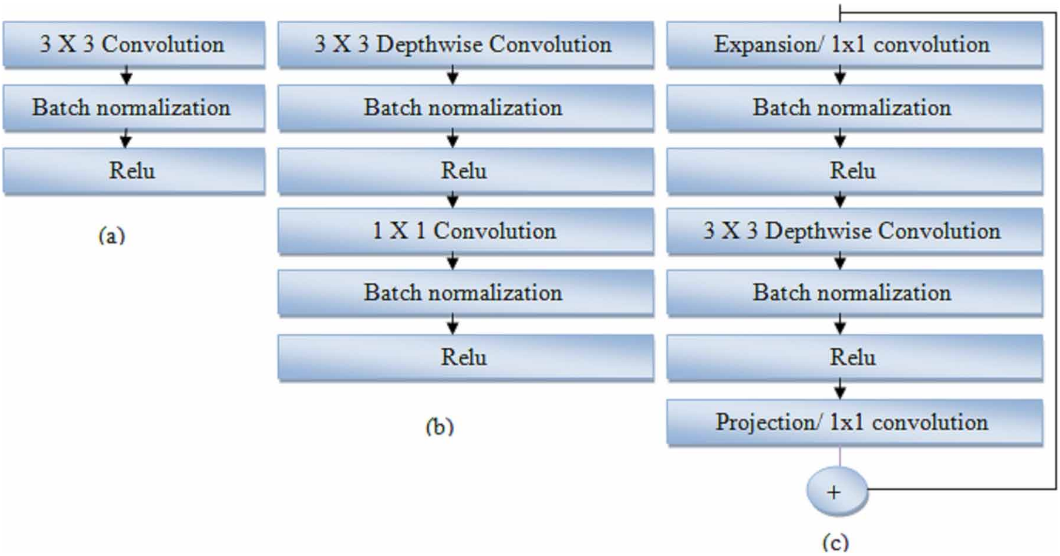
The backpropagation algorithm is used to update the network's parameters, the purpose of such method is to optimize the loss value based on various optimizers such as: the stochastic gradient descent (Robbins et al., 1951), Adam (Kingma et al., 2014), Rmsprop (Tieleman et al., 2017), momentum (Rumelhart et al., 1988) and Adagrad (Duchi et al., 2011).

The gradient descent (GD) optimizer updates the parameters according to Equation 2, where w is the parameter to update, ∇ is the gradient, L is the cost function and η is the learning rate. There are three variants of GD: the batch gradient descent (BGD), the stochastic gradient descent (SGD) and the mini-batch gradient descent (Ruder et al., 2016). In the BGD, the gradient of the whole dataset is computed at each update which can slow the training time for large datasets. Whereas in the SGD, the gradient is based on a single example and the computed gradient represents an approximation to the real one. The mini-batch gradient descent is designed to make a tradeoff between the exact behavior of BGD and the stochastic behavior of SGD, it considers only a portion of k training examples from the whole dataset to compute the gradient. To speed up the mini-batch learning, different methods have been proposed, such as momentum and RmsProp.

$$w^{(t+1)} = w^{(t)} - \eta \nabla_w L\left(w^{(t)}\right) \tag{2}$$

The RmsProp optimizer is a modified version of Rprop (Riedmiller et al., 1993). It combines between Rprop and SGD by dividing the learning rate by an exponentially decaying average of the

Figure 1. The structure of (a) standard convolutions, (b) depthwise separable convolutions (Howard et al., 2017) and (c) inverted residuals and linear bottlenecks structures (Sandler et al., 2018)



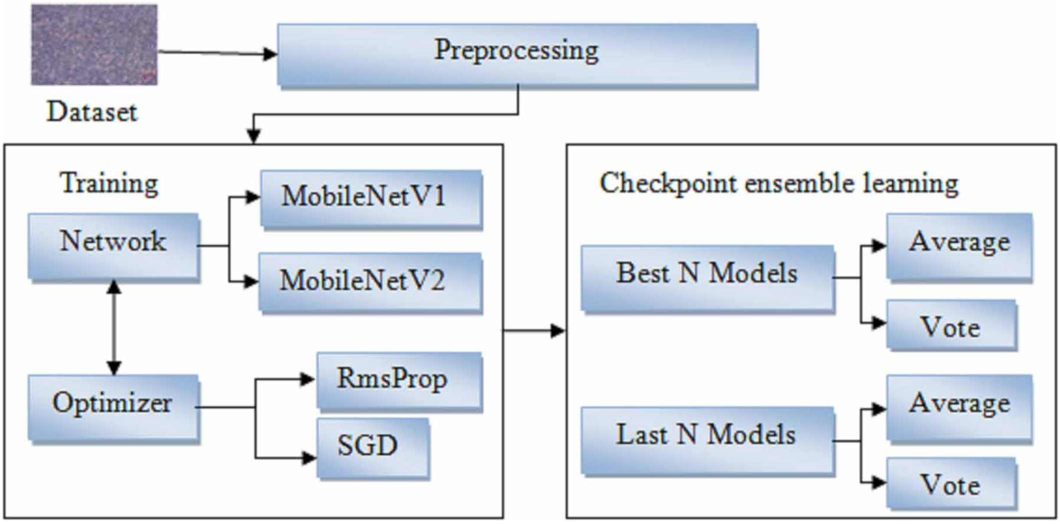
squared gradients (Tieleman et al., 2017). Equation 3 illustrates the RmsProp process, where g is the gradient root mean square, and α is the decay rate.

$$\begin{cases} g^{(t)} = \alpha g^{(t-1)} + (1-\alpha)\left(\nabla_w L\left(w^{(t)}\right)\right)^2 \\ w^{(t+1)} = w^{(t)} - \dfrac{\eta}{g^{(t)} + \epsilon} \nabla_w L\left(w^{(t)}\right) \end{cases} \tag{3}$$

4.THE PROPOSED FRAMEWORK

Figure 2 presents the proposed framework, which is based on three main modules: preprocessing, training and ensemble learning.

Figure 2. The proposed framework components



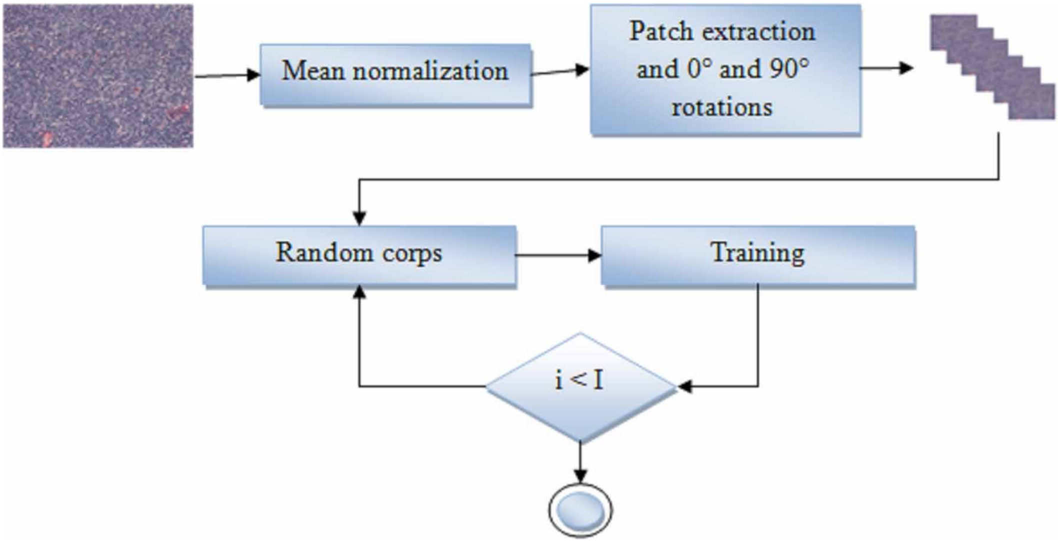
4.1. Preprocessing

Figure 3 illustrates the preprocessing scheme, which is based on mean normalization, patch extraction, rotations, and random corps.

The purpose of mean normalization is to center the data around zero means by subtracting from each image the mean of all images. Equation 4 presents the mean normalization process, where $m_{i,j}$ is the pixel value of the image matrix m and N is the number of images.

$$m_{i,j}(t) = m_{i,j}(t-1) - \frac{\sum_{k=1}^{N} m_{i,j}^{k}(t-1)}{N} \tag{4}$$

Figure 3. The preprocessing scheme



The resulting normalized images are then augmented by patch extraction (Figure 4) and rotations, this step seeks to take advantage of the high resolution of histopathological images to overcome overfitting on the small volumes of data. In this study, each image was divided into 144 x 144 non overlapped patches, then, the resulting patches were rotated by 0 and 90 degrees. These rotations contribute to improve the image analysis process since the pathologist can observe the biopsies from different angles. Finally, random samples of 128px x128px are extracted during training to improve the generalization.

4.2. Training and Ensemble Learning

Time and memory usage are critical for real deployments of computer vision systems, which present major challenges for convolutional neural networks (Table 4), moreover, these algorithms are prone to overfitting on small volumes of data. A reasonable approach to tackle these issues is the exploitation of small models since they have less overfitting problems and memory requirements (Howard et al., 2017). Thus, we selected the MobileNet architecture, due to its small number of parameters compared to other CNN architectures. This architecture is based on the depth-wise separable convolutions modules, that help to create very small image classification models.

Table 5 shows the number of parameters according to the input size (128 × 128) in the proposed framework.

Accurate models are characterized by their capacity to perform well on unseen samples during training and optimization. Therefore, to prevent overfitting, another used strategy was to select models according to their generalization efficiency. Our method was to follow up the learning process and to measure the gap between the accuracy curves of train, validation and test sets, where we used different types of optimizes: SGD with a variable learning rate, and the RmsProp optimizer. Then, test images are classified base on a vote between the classes of their sub-patches according to algorithm 1 (C is the number of classes, and P is the number of patches, if the t^{th} patch is classified with the class j then $d_{t,j} = 1$ else $d_{t,j} = 0$.

Finally, we proposed to combine the resulting models by an ensemble learning strategy to reduce their high variance. Usually, ensemble learning methods are characterized by three steps: model generation, model selection, and combination. In this research we used the checkpoint ensemble

Figure 4. The extracted patches from a CLL digitized image

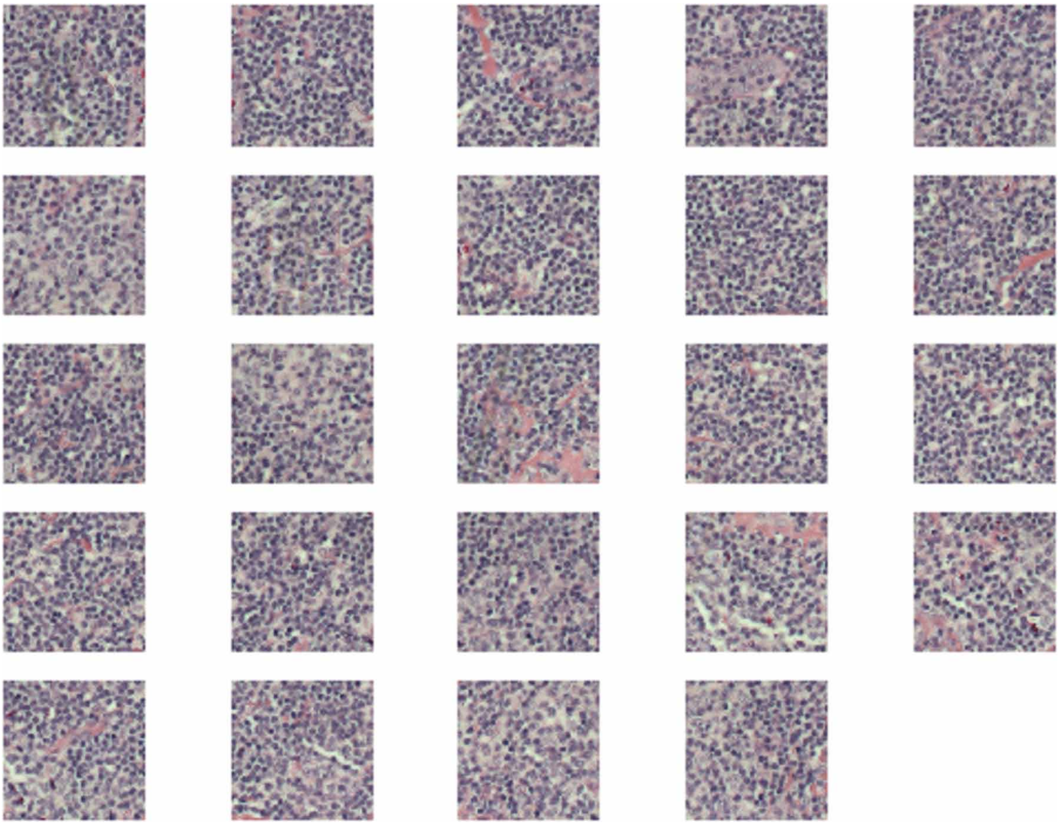


Table 4. The number of parameters and layers in CNN's architectures

Architecture	Number of parameters (Million)	Number of layers
AlexNet (Krizhevsky et al., 2012)	60	8
VGGNET (Simonyan et al., 2014)	133 to 144	11-19
Inception (Szegedy et al., 2015)	6.8	22
ResNet (He et al., 2016)	21.8 to 60.2 (Boulch, 2017)	18-152
ShuffleNet (Zhang et al., 2018)	3.4	50
(Howard et al., 2017)	4.2	28
(Sandler et al., 2018)	3.4	20

Table 5. The number of parameters in the used MobileNet architectures

Parameters	MobileNetV1	MobileNetV2
Total	3 231 939	2 261 827
Trainable	3 210 051	2 227 715
Non-trainable	21 888	34 112

method (algorithm 2), this technique requires only one training to generate N models instead of N training, which optimizes the training run time complexity. First, the set of models was generated by saving models at 3-minute intervals. Then, we employed two static ensemble selection strategies: the most accurate N models on the validation set and the last N saved models. Finally, voting and unweighted averaging methods were used to combine the selected models. Equation 5 illustrates the majority of voting process, where J is the selected class, C is the number of classes and N is the number of models if the t^{th} classifier chooses the class j then $d_{t,j} = 1$ else $d_{t,j} = 0$. Equation 6 presents the unweighted average process, where $y_t\left(x\right)$ is the weights vector of the t^{th} classifier for the sample x.

$$J = argmax_{j\in\left\{1,2,..,c\right\}} \sum_{t=1}^{N} d_{t,j} \tag{5}$$

$$y\left(x\right) = \frac{\sum_{t=1}^{N} y_t\left(x\right)}{N} \tag{6}$$

```
Algorithm 1. The testing process
Inputs: Test, Model.
Output: Test Accuracy.
Begin
Patches = Φ
for each (Image Test) do
Classes = Φ
Patches = Patch Extraction(Image)
for each(Patch ϵ Patches) do
Classes = Classes ∪ Model.Classify(Patch)
$Class = argmax_{j\epsilon \left\{1,2,..,c\right\}} \sum_{t=1}^{P} d_{t,j}$
Test _Accuracy = Compute_Test_Accuracy(Test)
End.
```

```
Algorithm 2. The combination process between the MobileNet
checkpoints
Inputs: Models, Validation, Test, Selection ∈ {Best, Last},
Combination ∈ {Vote, Average}, N: number of selected models.
Output: Test Accuracy.
Begin
if(Selection == Best) then
Models = Descending_Sort(Models, Validation)
Sub_Models = Models[1,N]
if (Combination == Vote) then
Test_Accuracy = vote(Sub_Models, Test)
else
Test_Accuracy = average(Sub_Models, Test)
End.
```

5. EXPERIMENTAL STUDY

In this study, the proposed approach was evaluated on the lymphoma histopathological dataset (Shamir et al., 2008), which was collected from 30 histological stained slides of lymph nodes. First, these slides were digitized by the Zeiss Axioscope light microscope and the AXio Cam MR5 camera. Then, high-resolution regions of interest have been extracted from the digitized images (1388 × 1040). Table 6 describes the number of images/patches in each sub-category (CLL, FL, MCL), where the different patches have been extracted from each image based on the data augmentation techniques described in the previous section.

For training, we employed both MobileNetV1 and MobileNetV2 networks, and two types of optimizers (Stochastic gradient descent and Rmsprop). Then, the stratified hold out method have been used to evaluate the trained networks: 20% for test, 20% of the rest of images for validation and 80% for training.

Table 7 presents the used parameters for training, where we employed a variable learning rate, with exponential decay for the SGD optimizer. The purpose of this variation is to prevent the slow convergence for small learning rate values and to overcome the oscillation problem in case of high learning rate values. The batch size was selected according to the memory requirements, and the loss value was computed based on the categorical cross-entropy loss function. This function is designed for the multi-label classification problems (Equation 7: S_p is the positive class score and C is the number of classes).

$$CE = -\log\left(\frac{e^{S_p}}{\sum_{j}^{C} x^{S_j}}\right) \tag{7}$$

Table 6. The Number of images/patches in the lymphoma dataset

Total number of images/ patches	Number of images/ patches by category		
375/41860	CLL	FL	MCL
	113/12600	140/15680	122/13580

The first set of analyses compares transfer learning and training from scratch techniques. The transfer learning process was carried out based on the transferred weights from the mobileNetV1 ImageNet model and the softmax classifier. To adapt the input images size to the ImageNet model requirements, we used two strategies: resize the high-resolution images to 224 x 224 and extract 224 x 224 sub-patches. The results, as shown in Table 8, indicate the efficiency of the patch-wise classification strategy, which is explained by the limitations of resizing.

Figure 5 highlights the cross-entropy convergence loss of the mobileNetV1 associated with RmsProp (a) and SGD (b) optimizers. Comparing the two curves, it can be seen that that SGD (loss = 0.3) has a fast convergence behavior compared to RmsProp (loss = 0.054).

Figure 6 shows the accuracy curves of models on: train, validation and test sets: (a) MobileNetV1 with RmsProp, (b) MobileNetV1 with SGD, (c) MobileNetV1 with SGD and dropout and (d) MobileNetV2 with SGD.

The MobileNetV1+ RmsProp curve highlights an oscillation and a high variance between the train and the test/validation accuracies. Whereas, MobileNetV1+ SGD reports an important correlation

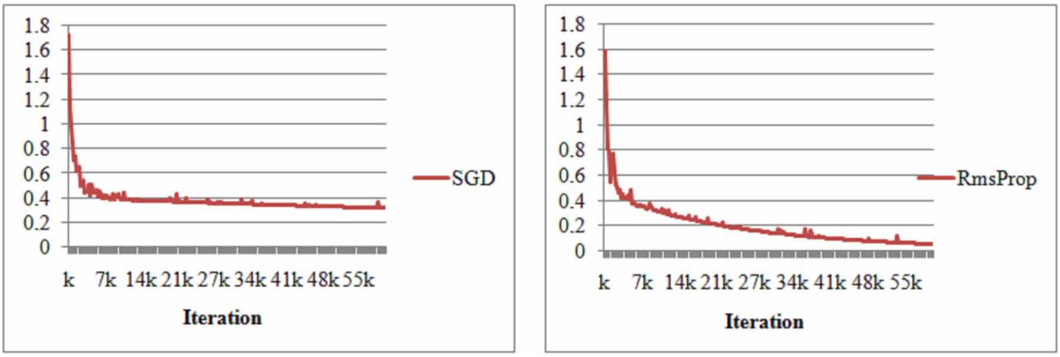
Table 7. The experimental study parameters

Parameter		Value
Maximum number of steps		60 000
Batch size (train/evaluation)		128/100
Learning rate (lr) (exponential decay)	Initial lr	0.05
	Lr decay	0.9
	Final lr	10^{-4} ×lr
RmsProp	Momentum	0.9
	Epsilon	0.9
Dropout rate		1.0
Loss function		Cross entropy
Dropout rate		0.2

Table 8. Transfer learning results

Patch size	Test accuracy
Resize Full Image	0.6533
224 x 224 Patches	0.7333

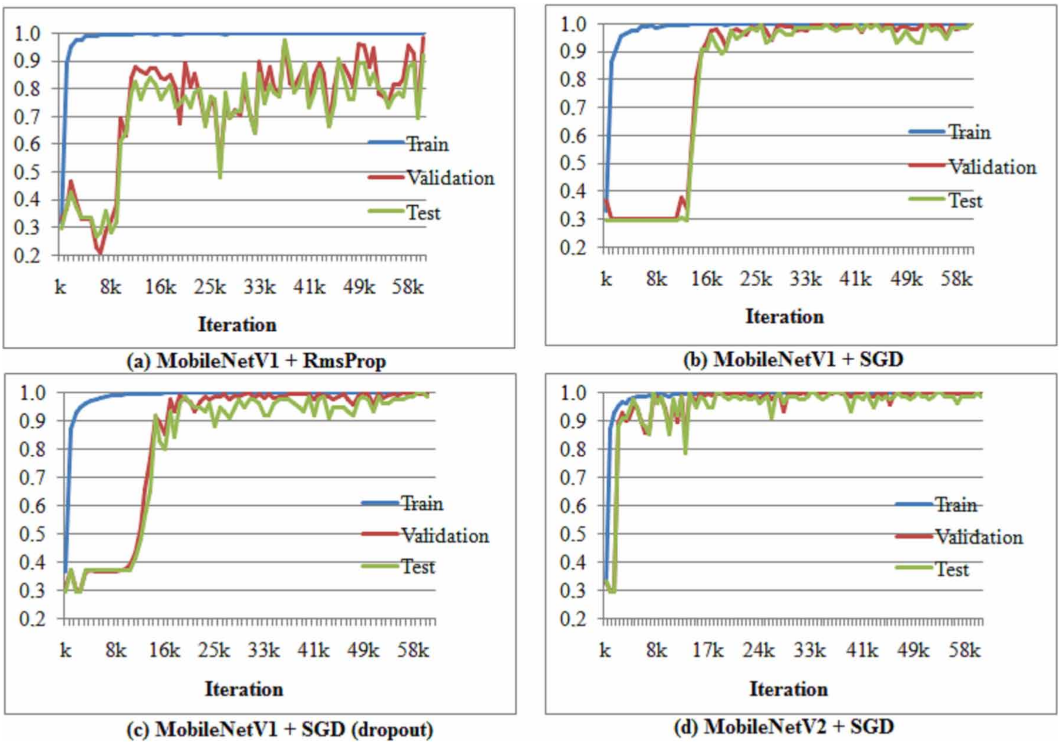
Figure 5. MobileNet loss curves based on RmsProp and SGD optimizers Train and validation loss curves



between the training and the validation/test curves starting from the 20k iteration, which provides an important insight on the capacity of the generated model in terms of generalization. Increasing the dropout rate from 0.01 to 0.2 confirmed this hypothesis, where the generated curves are similar in the case of the two dropout rate values. Finally, when comparing the two results on MobileNetV1 and MobileNetV2, it can be seen that MobileNetV2 provides more generalization on the test set.

For inference, we selected from the set of checkpoint models the model that maximizes the validation accuracy. Table 9 reports the obtained results by the selected model, where we observe that the difference between the train and validation sets accuracies was not significant. Whereas, it reveals that there has been a slight decrease for the test set due to the adopted voting strategy to

Figure 6. Train, validation and test accuracy curves of MobileNetV1 and MobileNetV2 models



(a) MobileNetV1 + RmsProp

(b) MobileNetV1 + SGD

(c) MobileNetV1 + SGD (dropout)

(d) MobileNetV2 + SGD

evaluate the wall test image. Taken together, these results indicate the efficiency of training from scratch compared to transfer learning for lymphomas subtypes classification.

Turning now to the experimental study based on the 5-cross-validation evaluation method and the MobileNetV2. The evaluation method splits the dataset into five groups and the model is trained five times on each group to generate five different models. For prediction, the results of models are averaged. The purpose of this method is to generate more confident models compared to a single evaluation, also it helps to improve the generalization ability of the classification framework and to overcome different problems like overfitting. The obtained results are illustrated in Table 10. The findings report a significant difference between the obtained results by the models 2 and 3, which provide further support for this evaluation method to reduce the high variance between neural networks results. The false generated predictions by the model 2 highlight its confusion between the MCL and the CLL classes, where two MCL and two CLL images have been misclassified.

Table 9. The obtained results by MobilenetV1 and MobileNetV2 on the train, validation and test sets (accuracy)

Network	Optimizer	Train	Validation	Test
MobilenetV1	RmsProp	0.9995	0.9866	0.9200
	SGD	0.9998	0.9990	1
	SGD (dropout)	0.9996	0.9981	0.9733
MobileNetV2	SGD	0.9996	0.9996	0.9867

Table 10. The obtained results by the MobileNetV2 based on the five cross-validation evaluation method

Division	1	2	3	4	5	Average
Validation	0.9996	0.9981	0.9977	0.9983	0.9987	0.9985
Test	0.9867	0.9467	1	1	0.9600	0.9787

In the final part of the experiments, we explored the ensemble checkpoint method, where two static ensemble selection strategies have been used: N best models and N last checkpoint models. Then, the selected models have been combined based on voting and averaging methods. Table 11 illustrates the obtained results on the test set, where the number of checkpoint models is 2 to 50 models. The findings reported here suggest that 3 to 4 models are enough to enhance the accuracy from 97.87% to 98.67% for the two strategies. Whereas a loss of accuracy was detected when combining 15 to 20 best models, which reveals that there was no clear evidence on the impact of the number of models on the results, as the loss of accuracy is related to the presence of week learners. The obtained results by the averaging method show that combining the last checkpoints tends to perform better than combining the best models, it can thus be related to the lack of diversity between the best models.

In summary, the findings reported here suggest the efficiency of the SDG compared to the Rmsprop optimizer for the MobileNet network and the lymphomas classification task. Moreover, these results provide further support for the hypothesis on the efficiency of checkpoint ensemble methods in deep learning applications. This study has been one of the first attempts to thoroughly examine the checkpoint ensembling for histopathological images classification.

Table 12 compares the obtained and the literature results on the lymphoma dataset. The best obtained results in the machine learning category have been achieved based on the segmented images with GA (Tosta et al., 2018) and the extracted visual texture descriptors (Song et al., 2016). Whereas in the proposed hybrid methods between machine and deep learning strategies, we observe that the hybridization between local features and the extracted features from pre-trained CNN models achieved good results compared to some machine learning methods (Shamir et al., 2008; Meng et al., 2010; Nava et al., 2016). The final part highlights the results of the deep learning methods, where we observe the efficiency of the MobileNetV2 compared to the AlexNet network. Moreover, these results reveal that deeper networks such as ResNet50, DenseNet were less promising. On the other hand, the comparative study between InceptionV3 and MobileNet reveals that results are close. Overall, this study reveals the importance of MobileNets compared to other deeper networks such as ResNet50, DenseNet, and InceptionV3 in terms of precision and computational complexity for lymphomas subtypes classification.

Table 11. The obtained results by the ensemble checkpoint method based on the last and the best combination strategies

Models	Method	Best models	Last checkpoints	Method	Best models	Last checkpoints
2	*Vote*	0.9813	0.9786	Average	0.9840	0.9813
3		0.9867	0.9813		0.9840	0.9867
4		0.9867	0.9867		0.9840	0.9813
5		0.9867	0.9840		0.9840	0.9867
10		0.9867	0.9840		0.9813	0.9840
15		0.9840	0.9867		0.9840	0.9840
20		0.9840	0.9867		0.9840	0.9840
50		0.9867	0.9867		0.9813	0.9813

Table 12. Comparison between the obtained and the literature results

Method	Reference	Evaluation Method	Accuracy (%)	
Machine learning	(Shamir et al., 2008)	-	85	
	(Meng et al., 2010)	3-cross-validation	92.70	
	(Nava et al., 2016)	10-cross-validation	93.83	
	(Song et al., 2016)	5- cross-validation	96.8	
	(Di Ruberto et al., 2015)	cross-validation	96.4	
	(Tosta et al., 2018)	10-cross-validation	98.14	
	(Orlov et al., 2010)	8-cross-validation	98-99	
Machine + Deep learning	(Codella et al., 2016)	3-cross-validation	95.5	
	(Song et al., 2017)	4-cross-validation	96.5	
	(Nanni et al., 2018)	-	97.33	
	(Song et al., 2017b)	4-cross-validation	97.9	
	(Bai et al., 2019)	hold out	99.1	
	(Nanni et al., 2019)	5-cross-validation	96.87	
Deep learning	(Janowczyk et al., 2016)	5-cross-validation	96.58	
	(Nanni (b) et al., 2019)	5-cross-validation	ResNet50	92.00
			DenseNet	93.60
	Ours	Hold-out	Inception-v3	97.78
	Ours (MobileNetV2)	5-cross-validation	97.87	
	Ours (ensemble MobileNetV2)	5-cross-validation	98.67	

6. CONCLUSION

In this research, we presented a deep neural network framework based on MobileNet architecture. The main purpose of this study was to use a maximum of regularization techniques to overcome overfitting: data augmentation, small models, K-cross-validation and ensemble learning. First, the train set has been augmented by data augmentation techniques. Then, the appropriate model in terms of generalization has been selected from the set of generated models by the combination of two MobileNets architectures with two types of optimizers. Finally, we combined checkpoint models by voting and averaging. The proposed framework was tested on the lymphoma dataset, and the best results have been obtained by the voting method between the 3 best checkpoints. The comparative study with the literature results provides important insights into the benefit of checkpoint ensembling in computer-aided diagnostic systems. However, the averaging and voting methods are sensitive to week learners, which can produce a loss in accuracy.

As perspective to this work, we are looking to work on other more challenging histopathological datasets such as BreakHis and the ovarian carcinoma dataset and other ensemble learning methods such as meta-learners.

6.1. Compliance With Ethical Standards

The authors declare that they have no conflict of interest. This article does not contain any studies with human participants performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

ACKNOWLEDGMENT

This work was sponsored by the Dgrsdt.

REFERENCES

Al-Janabi, S., Huisman, A., & Van Diest, P. J. (2012). Digital pathology: Current status and future perspectives. *Histopathology*, *61*(1), 1–9. doi:10.1111/j.1365-2559.2011.03814.x PMID:21477260

Bai, J., Jiang, H., Li, S., & Ma, X. (2019). NHL Pathological Image Classification Based on Hierarchical Local Information and GoogLeNet-Based Representations. *BioMed Research International*. PMID:31016181

Boulch, A. (2017). Sharesnet: reducing residual network parameter number by sharing weights.

Chan, J. K. (2001). The new World Health Organization classification of lymphomas: The past, the present and the future. *Hematological Oncology*, *19*(4), 129–150. doi:10.1002/hon.660 PMID:11754390

Chen, H., Dou, Q., Wang, X., Qin, J., & Heng, P. A. (2016, February). Mitosis detection in breast cancer histology images via deep cascaded networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* AAAI Press.

Chen, H., Lundberg, S., & Lee, S. I. (2017). Checkpoint Ensembles: Ensemble Methods from a Single Training Process.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258). IEEE Press. doi:10.1109/CVPR.2017.195

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013, September). Mitosis detection in breast cancer histology images with deep neural networks. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 411-418). Springer. doi:10.1007/978-3-642-40763-5_51

Codella, N., Moradi, M., Matasar, M., Sveda-Mahmood, T., & Smith, J. R. (2016, March). Lymphoma diagnosis in histopathology using a multi-stage visual learning approach. In Medical Imaging 2016: Digital Pathology. International Society for Optics and Photonics.

Deng, L., & Platt, J. C. (2014). Ensemble deep learning for speech recognition. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*.

Di Ruberto, C., Fodde, G., & Putzu, L. (2015, September). On different colour spaces for medical colour image classification. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns* (pp. 477-488). Springer. doi:10.1007/978-3-319-23192-1_40

Dif, N., & Elberrichi, Z. (2019). An Enhanced Recursive Firefly Algorithm for Informative Gene Selection. *International Journal of Swarm Intelligence Research*, *10*(2), 21–33. doi:10.4018/IJSIR.2019040102

Dif, N., & Elberrichi, Z. (2018). A Multi-Verse Optimizer Approach for Instance Selection and Optimizing 1-NN Algorithm. *International Journal of Strategic Information Technology and Applications*, *9*(2), 35–49. doi:10.4018/IJSITA.2018040103

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

Fok, W., Jamart, K., Zhao, J., & Fernandez, J. (2018, September). Ensemble of Convolutional Neural Networks for Heart Segmentation. In *Proceedings of the International Workshop on Statistical Atlases and Computational Models of the Heart* (pp. 282-291). Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE Press.

Hervé, N., Servais, A., Thervet, E., Olivo-Marin, J. C., & Meas-Yedid, V. (2011, March). Statistical color texture descriptors for histological images analysis. In *Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (pp. 724-727). IEEE. doi:10.1109/ISBI.2011.5872508

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7. PMID:27563488

Ju, C., Bibaut, A., & van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, *45*(15), 2800–2818. doi:10.1080/02664763.2018.1441383 PMID:31631918

Jung, H., Kim, B., Lee, I., Lee, J., & Kang, J. (2018). Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method. *BMC Medical Imaging*, *18*(1), 48. doi:10.1186/s12880-018-0286-0 PMID:30509191

Kainz, P., Pfeiffer, M., & Urschler, M. (2015). Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization.

Kobayashi, H. (2018). Frustratingly Easy Model Ensemble for Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4165-4176). Academic Press. doi:10.18653/v1/D18-1449

Komura, D., & Ishikawa, S. (2018). Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, *16*, 34–42. doi:10.1016/j.csbj.2018.01.001 PMID:30275936

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105). Academic Press.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & Sánchez, C. I. et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. doi:10.1016/j.media.2017.07.005 PMID:28778026

Meng, T., Lin, L., Shyu, M. L., & Chen, S. C. (2010, December). *Histology image classification using supervised classification and multimodal fusion. In Proceedings of the 2010 IEEE international symposium on multimedia* (pp. 145–152). IEEE Press.

Mosca, A., & Magoulas, G. D. (2016, September). *Deep Incremental Boosting*. GCAI.

Nanni, L., Ghidoni, S., & Brahnam, S. (2018). *Ensemble of convolutional neural networks for bioimage classification*. Applied Computing and Informatics. doi:10.1016/j.aci.2018.06.002

Nanni, L., Brahnam, S., Ghidoni, S., & Maguolo, G. (2019). *General Purpose (GenP)*. Bioimage Ensemble of Handcrafted and Learned Features with Data Augmentation.

Nanni (b), L., Brahnam, S., & Maguolo, G. (2019). Data Augmentation for Building an Ensemble of Convolutional Neural Networks. In *Innovation in Medicine and Healthcare Systems, and Multimedia* (pp. 61-69). Springer.

Nava, R., González, G., Kybic, J., & Escalante-Ramírez, B. (2016, December). Characterization of hematologic malignancies based on discrete orthogonal moments. In *Proceedings of the 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE. doi:10.1109/IPTA.2016.7821039

Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 443-449). ACM. doi:10.1145/2818346.2830593

Orlov, N. V., Chen, W. W., Eckley, D. M., Macura, T. J., Shamir, L., Jaffe, E. S., & Goldberg, I. G. (2010). Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, *14*(4), 1003–1013. doi:10.1109/TITB.2010.2050695 PMID:20659835

Riedmiller, M., & Braun, H. (1993, March). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proceedings of the IEEE international conference on neural networks* (pp. 586-591). IEEE Press. doi:10.1109/ICNN.1993.298623

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*(3), 400–407. doi:10.1214/aoms/1177729586

Roberto, G. F., Neves, L. A., Nascimento, M. Z., Tosta, T. A., Longo, L. C., Martins, A. S., & Faria, P. R. (2017). Features based on the percolation theory for quantification of non-Hodgkin lymphomas. *Computers in Biology and Medicine*, *91*, 135–147. doi:10.1016/j.compbiomed.2017.10.012 PMID:29059591

Ruder, S. (2016). An overview of gradient descent optimization algorithms.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling, 5*(3), 1.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510-4520). IEEE Press.

Sang, D. V., Cuong, D. M., & Cuong, L. T. B. (2018, December). An Effective Ensemble Deep Learning Framework for Malware Detection. In *Proceedings of the Ninth International Symposium on Information and Communication Technology* (pp. 192-199). ACM. doi:10.1145/3287921.3287971

Sang, D. V., & Ha, P. T. (2018, April). Discriminative deep feature learning for facial emotion recognition. In *Proceedings of the 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)* (pp. 1-6). IEEE. doi:10.1109/MAPR.2018.8337514

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T. J., & Goldberg, I. G. (2008). IICBU 2008: A proposed benchmark suite for biological image analysis. *Medical & Biological Engineering & Computing*, *46*(9), 943–947. doi:10.1007/s11517-008-0380-5 PMID:18668273

Shankland, K. R., Armitage, J. O., & Hancock, B. W. (2012). Non-hodgkin lymphoma. *Lancet*, *380*(9844), 848–857. doi:10.1016/S0140-6736(12)60605-9 PMID:22835603

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Skalic, M., Pekalski, M., & Pan, X. E. (2017). Deep learning methods for efficient large scale video labeling.

Song, Y., Cai, W., Huang, H., Feng, D., Wang, Y., & Chen, M. (2016). Bioimage classification with subcategory discriminant transform of high dimensional visual descriptors. *BMC Bioinformatics*, *17*(1), 465. doi:10.1186/s12859-016-1318-9 PMID:27852213

Song, Y., Chang, H., Huang, H., & Cai, W. (2017, September). Supervised intra-embedding of fisher vectors for histopathology image classification. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 99-106). Springer. doi:10.1007/978-3-319-66179-7_12

Song, Y., Li, Q., Huang, H., Feng, D., Chen, M., & Cai, W. (2017b). Low dimensional representation of fisher vectors for microscopy image classification. *IEEE Transactions on Medical Imaging*, *36*(8), 1636–1649.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). IEEE Press.

Tieleman, T., & Hinton, G. (2017). *Divide the gradient by a running average of its recent magnitude*. Coursera.

Tosta, T. A. A., Neves, L. A., & do Nascimento, M. Z. (2017). Segmentation methods of H&E-stained histological images of lymphoma: a review. *Informatics in medicine unlocked, 9*, 35-43.

Tosta, T. A., de Faria, P. R., Neves, L. A., & do Nascimento, M. Z. (2018, April). Fitness Functions Evaluation for Segmentation of Lymphoma Histological Images Using Genetic Algorithm. In *Proceedings of the International Conference on the Applications of Evolutionary Computation* (pp. 47-62). Springer. doi:10.1007/978-3-319-77538-8_4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). Academic Press.

Vujasinovic, T., Pribic, J., Kanjer, K., Milosevic, N. T., Tomasevic, Z., Milovanovic, Z., & Radulovic, M. et al. (2015). Gray-level co-occurrence matrix texture analysis of breast tumor images in prognosis of distant metastasis risk. *Microscopy and Microanalysis*, *21*(3), 646–654. doi:10.1017/S1431927615000379 PMID:25857827

Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., . . . Madabhushi, A. (2014, March). Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. In Medical Imaging 2014: Digital Pathology (p. 90410B). International Society for Optics and Photonics.

Xu, J., Zhou, C., Lang, B., & Liu, Q. (2017). Deep learning for histopathological image analysis: Towards computerized diagnosis on cancers. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing* (pp. 73–95). Cham: Springer. doi:10.1007/978-3-319-42999-1_6

Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *Proceedings of the European conference on computer vision* (pp. 818-833). Springer.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6848-6856). IEEE Press. doi:10.1109/CVPR.2018.00716

Nassima Dif is a PhD student at Djillali Liabes University.

Zakaria Elberrichi is a professor and a team director.