Dealing with Lack of Training Data for Convolutional Neural Networks: The Case of Digital Pathology

*electronics*

MDPI

*Article*

# Dealing with Lack of Training Data for Convolutional Neural Networks: The Case of Digital Pathology

**Francesco Ponzio, Gianvito Urgese, Elisa Ficarra and Santa Di Cataldo ***

Department of Control and Computer Engineering, Politecnico di Torino, Cso Duca degli Abruzzi 24, 10129 Torino, Italy; francesco.ponzio@polito.it (F.P.); gianvito.urgese@polito.it (G.U.); elisa.ficarra@polito.it (E.F.)
* Correspondence: santa.dicataldo@polito.it; Tel.: +39-011-090-7020

Received: 25 January 2019; Accepted: 21 February 2019; Published: 26 February 2019


**Abstract:** Thanks to their capability to learn generalizable descriptors directly from images, deep Convolutional Neural Networks (CNNs) seem the ideal solution to most pattern recognition problems. On the other hand, to learn the image representation, CNNs need huge sets of annotated samples that are unfeasible in many every-day scenarios. This is the case, for example, of Computer-Aided Diagnosis (CAD) systems for digital pathology, where additional challenges are posed by the high variability of the cancerous tissue characteristics. In our experiments, state-of-the-art CNNs trained from scratch on histological images were less accurate and less robust to variability than a traditional machine learning framework, highlighting all the issues of fully training deep networks with limited data from real patients. To solve this problem, we designed and compared three transfer learning frameworks, leveraging CNNs pre-trained on non-medical images. This approach obtained very high accuracy, requiring much less computational resource for the training. Our findings demonstrate that transfer learning is a solution to the automated classification of histological samples and solves the problem of designing accurate and computationally-efficient CAD systems with limited training data.

**Keywords:** convolutional neural networks; deep learning; histological image analysis; computer-aided diagnosis systems; transfer learning

## 1. Introduction

Histological image analysis is the gold standard for the primary diagnosis and assessment of c large number of cancers [1]. Typically, when a cancer is suspected, the patient undergoes a biopsy, and a thin layer of sample tissue is resected and mounted on a slide after fixation and staining, for example by Hematoxylin and Eosin (H&E). Then, the slide is analyzed by a pathologist looking for possible alterations of the normal tissue architecture, categorized into a number of classes of interest.

The diffusion of digital scanners, able to transform the physical histological slides into multi-resolution digital resources called Whole-Slide Images (WSIs), is rapidly changing the workflow of clinical laboratories [2]. Traditional histopathology, based on visual evaluation of the samples directly under the microscope, is being progressively abandoned in favor of Computer-Aided Diagnosis (CAD) systems, fostering a complete automatization of downstream image analysis.

On the one hand, automated image analysis is a major improvement on human assessment, which has been majorly affected by inter- and intra-observer variability [3]. On the other hand, it is challenged by the size (in the order of gigapixels), as well as by the high complexity and variability of the histological images. The origin of such variability is three-fold: (i) "biological", due to different cells (either cancerous or not) and corpuscles of variable appearance normally coexisting in a specimen; (ii) "pathological", due to unpredictable alterations of the tissue architecture induced by the cancer; and (iii) "technological", due to inconsistent staining, as well as to a typical lack of standards in the

image generation and acquisition process. On these grounds, the design of automated algorithms for the accurate assessment of histological images is still a very open research problem.

The most consolidated systems for histopathology assessment typically rely on classic texture analysis [4]. Image texture provides information about the spatial arrangement of color or intensities in an image. Hence, when applied to histological images, it can be used to characterize the spatial arrangement of the cells or in general the architecture of a tissue [5]. Systems based on texture analysis, as the name suggests, leverage the extraction of a limited set of texture and morphometric descriptors from the histological images; for example, statistic descriptors based on the Grey-Level Co-occurrence Matrices (GLCM), Local Binary Patterns (LBPs) and its variants, features based on Gabor or wavelet transform, and key-point detectors and descriptors such as Speeded-Up Robust Features (SURF) [6,7]. The local features set, eventually encoded into compact dictionaries by means of clustering techniques, as in the Bag Of Features method (BOF) [8], is then fed into a classifier to predict the label of the input specimens [9].

The major issue of all the classical texture-analysis approaches is typically the dependency on a fixed set of handcrafted features for the image representation [4,5]. Indeed, pre-designing the features based on a priori assumptions on the patterns that are most important for the classification intrinsically limits the robustness and generalization capabilities of a system. In the specific case of histopathology, many challenges are posed by the high intra-class variability, as well as by a general lack of agreement among the medical community on the histological characterization of complex pathologies (especially in the case of rare or lesser-known forms of cancer). To overcome this issue, in this work, we address the problem of automated histological classification using a *feature learning* approach, where the expert system learns a set of discriminant features directly from the images without any a priori constraint on the image representation.

Among feature learning methods, deep learning and, more specifically, Convolutional Neural Networks (CNNs) have now become a major trend in many computer vision and medical tasks [10–12]. In CNNs, a number of convolutional and pooling layers learn by backpropagation the set of features that are best for classification, thus avoiding the design of handcrafted texture descriptors. Nonetheless, the necessity of training the networks with a huge number of independent annotated images (typically in the order of tens of thousands at least) is still an open issue, which limits their usability in the everyday clinical setting.

*Transfer learning* (i.e., using CNNs pre-trained on different types of images, for which large datasets are available) seems a good solution to this problem, but only on the condition that the transfer happens between two similar imaging domains [13]. Only recently, cross-domain transfer learning has also been considered, with some promising results even in histological image analysis [14,15]. This opens the way to more in-depth research on the practical use of CNNs (and pre-trained CNNs in particular) in everyday histopathology, with typical problems being the limited availability of computational resources and annotated datasets.

In this work, we evaluate a CNN-based approach to perform the automated assessment of histological samples, targeting multi-class image characterization problems with H&E-stained WSIs as the input. More specifically, we seek an answer to the following research questions: (i) Are CNNs a good solution for histopathological image classification, as they are in other computer vision applications? (ii) How is it possible to cope in practice with the scarcity of annotated training samples? (iii) Is transfer learning a viable solution, and how should the transfer learning system be designed to boost the accuracy and generalizability of the results?

To answer such questions, we fully train several CNN models on histopathological images and assess their accuracy on an independent test set. This technique is experimentally compared with three different transfer learning approaches, leveraging on CNNs pre-trained on a dataset from a completely different context. The first transfer learning approach uses the pre-trained CNN to extract a set of discriminative features that will be fed into a separate support vector machines classifier. The second approach fine-tunes on histological images only a few stages of the pre-trained CNN. The third

approach uses the weights learned from a different context just to initialize the training. Finally, we perform a comprehensive comparative assessment and in-depth discussion of the transfer learning capabilities of CNNs in the domain of digital pathology. The final aim of our investigation is to provide a generalizable transfer learning framework, both in terms of architecture and training paradigm, that can be successfully applied to any other classification problem affected by intra-class variability and a lack of training data.

A very preliminary version of this study was recently presented in a conference paper, targeting the specific problem of colorectal image classification [16]. In the current paper, we address the problem of histological image classification from a general point of view, using colorectal polyps assessment, which is an important and challenging problem in histopathology and medicine, just as a case study. Further experiments on three additional datasets, targeting different histological categories and anatomical tissues (i.e. cardiovascular, bone, and pleural tissues), are also provided in order to prove the generality and robustness of our findings. On top of that, we extend our investigation to several CNN models, with different depth and architectural characteristics, and introduce an additional transfer learning approach to our study. Finally, we add a traditional machine learning framework, based on BOF encoding and the SVM classifier, as a reference for the accuracy assessment.

This paper is organized as follows. In Section 2, we characterize the main case study and the datasets used for our experiments. In Section 3, we introduce the CNN models and describe in detail the design and strategy of the histological classification approaches. In Section 4, we report our experimental results. In Section 5, we discuss our main findings. Finally, Section 6 concludes the paper.

## 2. Materials

In this section, we describe our main case study and characterize the datasets used in our experiments. All the image data are freely available from the authors on request.

### *2.1. Colorectal Polyps Assessment*

In our work, we chose the histopathological assessment of colorectal polyps as the main case study. This is a complex multi-class classification problem, where the categories are subject to a significant level of intra-class variability. Hence, it is a relevant case study for classification techniques that do not leverage on a fixed set of pre-designed features.

According to the World Health Organization, Colorectal Cancer (CRC) is the second most common tumor worldwide. While CRC is associated with high (around or above 90%) chances of 5-year relative survival when it is found at a very early stage, only around 40% of the colorectal polyps are found and removed before they eventually develop into malignant tumors. This has a tremendous impact on the mortality rate and makes CRC one of the leading causes of cancer-related death in most Western countries [3,17]. The issue is receiving considerable attention from the healthcare systems, which are now highly investing in mass-screening programs and diagnostic systems for CRC.

From a histological point of view, CRC generally originates from the most internal layer of the intestinal wall as an abnormal tissue growth called a polyp or *adenoma*, whose irregular neoplastic cells tend to infiltrate the other layers of the wall. During a colonoscopy, the physician visually inspects the lining of the colon and identifies the polyps, which are eventually resected via biopsy to undergo histopathological analysis.

A growth on the inner surface of the colon that does not show specific architectural irregularities compared to healthy tissue and carries a very low risk of developing a cancer is called a *hyperplastic polyp*. Adenomas, on the other hand, can show different types of tissue irregularities, which are associated with different types of precancerous growth. According to most of the literature on CRC, there are two major classes of precancerous colorectal polyps. The first class is *conventional adenomas*, which are the precursor of around 70% of all colorectal cancers. Depending on their architecture, conventional adenomas can be classified as either *tubular* or *villous* or tubulo-villous,

in case there is a mixture of the two elements. The second class of precancerous polyps, *serrated adenomas*, have a specific saw-tooth appearance and are currently understood to be the precursor of the remaining 30% of colorectal cancers, but from a pathway of genetic alterations that is different from the conventional adenomas.

In the light of the above, identifying and categorizing the polyps at a very early stage has important implications for successful detection, surveillance, and personalized treatment of different types of colorectal cancers [3,17].

Our previous work on colorectal image analysis targeted a simplified classification problem where adenomas were treated as a unicum, overlooking their specific sub-classes [16]. In this paper, we address the full histological assessment of colorectal polyps (malignant, non-malignant, and cancer precursors), considering five main histological categories: (i) Adenocarcinoma (AC), where there is evidence of a conclamated CRC; (ii) Hyper-plastic polyp (H); (iii) traditional Serrated adenoma (S); (iv) Tubular adenoma (T); and (v) Villous adenoma (V). Several examples of each category are shown in Figure 1.

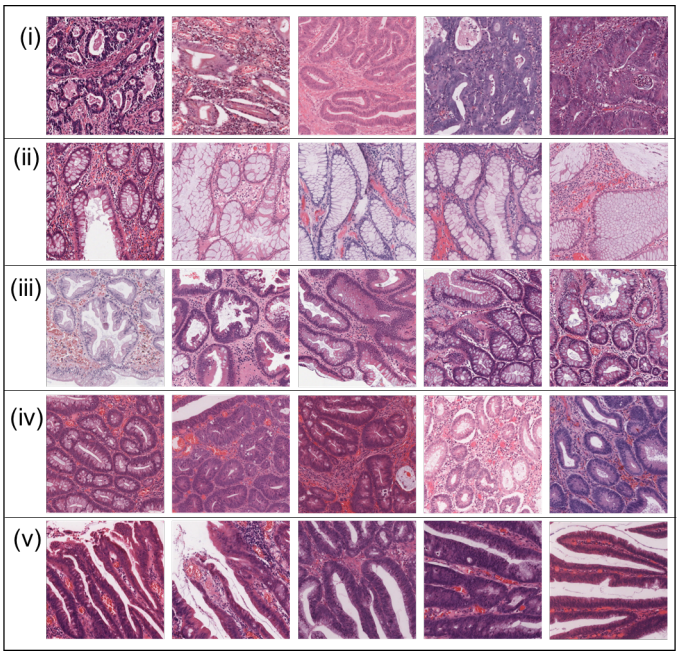


**Figure 1.** H&E samples of colorectal polyps. The figure shows five different histological categories of polyps, one per line: (**i**) Adenocarcinoma (AC); (**ii**) Hyperplastic polyp (H); (**iii**) Serrated adenoma (S); (**iv**) Tubular adenoma (T); (**v**) Villous adenoma (V).

For our study, we obtained 41 hematoxylin and eosin colon tissue slides from the Virtual Pathology Slide Library of the University of Leeds, a repository of histological samples that have been digitized and curated by a trained pathologist. The digitalized slides are stored with their anonymized clinical information and publicly available at http://www.virtualpathology.leeds.ac.uk/.

Each slide belongs to a single patient diagnosed with one of the five classes of lesions reported in Figure 1. The original data were in the form of WSIs (also known as virtual slides), the de-facto standard of the modern histological scanners. To facilitate efficient viewing via specialized vendor-supplied software, these very large files were encoded into multilayered pyramidal structures across multiple resolutions (e.g., $1\times$, $20\times$, $40\times$).

To make the histological images usable by convolutional neural networks, we cropped the WSIs into a large number of small square patches. As the characterization of the colonic polyps depends on both morphological and textural information (respectively, the shape and structure and the cytological characteristics of the glands), the choice of the magnification factor for the cropping may be crucial. More specifically, lower magnifications ensure a better view of tissue architecture,

while larger magnifications provide a better characterization of the micro-textural characteristics of the cells (see the two examples in Figure 2). On the other hand, the number of independent patches that can be cropped from the same WSI is proportional to the magnification factor.

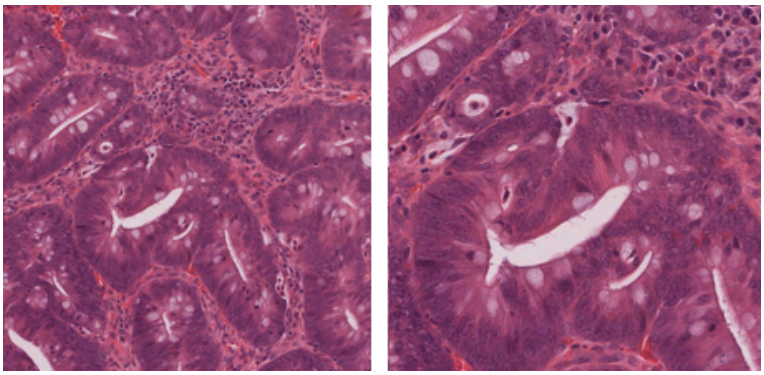

**Figure 2.** Colorectal tissue patches of a tubular adenoma at two different magnifications: 20× (**left**) and 40× (**right**).

For our experiments, we cropped the WSIs into non-overlapping patches at two different magnification factors: 20× (patch-size of 2178 × 2178 pixels) and 40× (1089 × 1089 pixels), as reported in Table 1, and we tried to establish which magnification factor provided the best compromise between macro- and micro-textural representation. For this purpose, we ran preliminary tests with either 20× and 40× patches, as well as with both the magnification factors put together, and we found that 20× ensured the best performance in terms of classification accuracy on the five targeted classes of colorectal polyps. Hence, for the sake of readability, in the following, we show and discuss only the results obtained on the 20× dataset.

For training and testing purposes, our dataset was divided into two disjointed sub-cohorts, comprised of 32 patients for training and 9 for testing (see Table 1).

**Table 1.** Colorectal image dataset. Characterization of the independent sets used for training and testing purposes.

| | | **Train** | **Test** | **Tot** |
|---|---|---|---|---|
| *# patches* | 20× | 10,052 | 2448 | 12,500 |
| | 40× | 15,876 | 4124 | 20000 |
| *# patients* | | 32 | 9 | 41 |

In order to avoid class imbalance, irrespective of the image magnification, both the training and test sets contained an almost equal number of patches of the five different categories. To compensate for possible color discrepancy, before being fed into the classifier, all the patches were normalized by the mean and standard deviation computed over the entirety of the training data.

### *2.2. Cardiovascular, Bone, and Pleural Tissue Assessment*

To prove the generality of our findings, we used three additional datasets, targeting different histological categories and anatomical tissues:

1. a *cardiovascular tissue dataset*, with five different histological categories: loose Connective tissue (CN), smooth muscle of Muscular Artery (MA), smooth muscle of the large Vein (VE), smooth muscle of the Elastic Artery (EA), and Cardiac muscle of the heart (HE). Images and corresponding annotations were obtained from [18].
2. a *bone tissue dataset*, again with five categories of interest: T Cells (TC), Osteoclasts (OS), Hydroxyapatite (HD), Parenchyma (PA), and regions with no clinical interest (VT). Images and annotations were obtained from a pathologist.

3. a *pleural tissue dataset*, with two categories of interest: Epithelioid Mesothelioma (EM) and Sarcomatoid Mesothelioma (SM). Images and annotations were obtained from a pathologist.

As can be gathered from the examples in Figure 3, the architectural characteristics of the first two datasets were especially different from our main case study (i.e., colorectal tissue), as well as from each other. Furthermore, while the cardiovascular dataset included classical H&E stained histological images, the tissue bone dataset was obtained with a immunohistochemical protocol, using FAST RED, FAST BLUE, and DAB chromophores for the staining. The third dataset was the most similar to the colorectal one from a histological point of view, as they both included epithelial tissue samples, but from a completely different anatomical location (i.e., lung pleura).

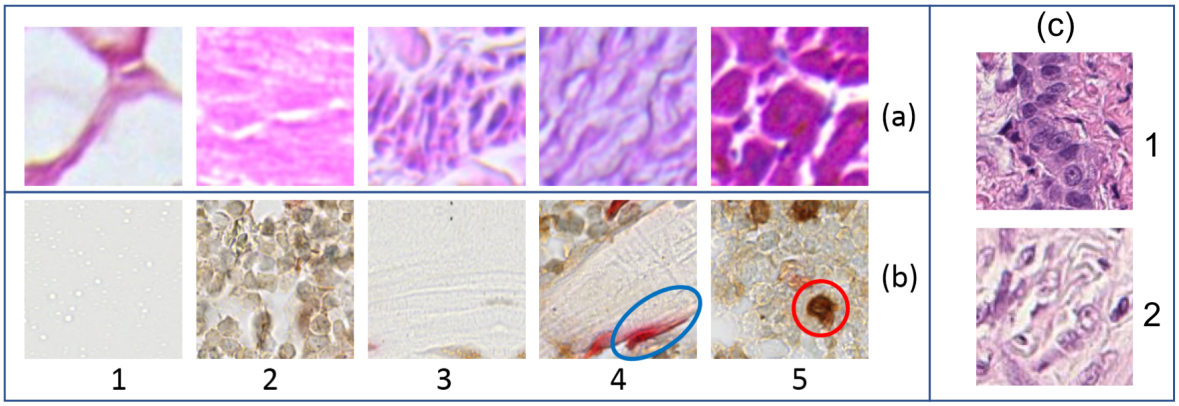


**Figure 3.** Additional histological datasets (examples of annotated patches). (**a**) *Cardiovascular tissue dataset*: (1) loose connective tissue, (2) smooth muscle of muscular artery, (3) smooth muscle of the large vein, (4) smooth muscle of the elastic artery, and (5) cardiac muscle of the heart. (**b**) *Bone tissue dataset*: (1) regions with no clinical interest, (2) parenchyma, (3) hydroxyapatite, (4) osteoclasts (the cell of interest is circled in blue), and (5) T cells (the cell of interest is circled in red). (**c**) *Pleural tissue dataset*: (1) epithelioid mesothelioma and (2) sarcomatoid mesothelioma.

All datasets consisted of 5000 annotated patches, equally balanced among the available classes. Eighty percent of the available patches were used for training and validation purposes and the remaining 20% for testing.

## 3. Methods

Convolutional Neural Networks (CNNs) are a class of feed-forward neural networks that have become in a very short time a major trend in most of the computer vision and pattern recognition applications. CNNs belong to the category of *deep networks*, where the *depth* is given by the presence of a high number of hidden stages compared to regular neural networks. More specifically, a CNN contains two different types of trainable stages:

1. a large number of locally-connected layers (the higher the number, the larger the depth of the CNN), devoted to learning the image representation. The features are learned on a hierarchical basis, with the first layers typically learning low-level features (e.g., simple edges) and successive layers learning features at a progressively-increasing level of abstraction (e.g., complex patterns and objects).
2. a small number of fully-connected layers at the end of the network, devoted to learning the classification task and basically acting like a traditional multilayer perceptron.

The possibility of combining a large depth with local connectivity allows learning a comprehensive set of image descriptors with a relatively low number of parameters to be learned.

From a computational point of view, the locally-connected stages of a CNN have two main building blocks:

1. *Convolutional (CONV)* blocks, which perform 2D convolution operations and eventually apply a non-linear activation function (for example, a Rectified Linear Unit (ReLU)) on the input image. Based on the trainable parameters of the kernels, the blocks can detect different types of local patterns.
2. *Pooling (POOL)* blocks, which perform a non-linear down-sampling of the input (for example, by max or mean functions). This has the double effect of reducing the number of parameters of the network, hence reducing the risk of overfitting, and of making the image representation spatially invariant.

While the CNN models like the LeNet consisted of simple regular sequences of very few CONV and POOL layers [19], the progressively wider availability of inexpensive computational resources and hardware acceleration has enabled over the years the design of much deeper models, such as AlexNet [20] and VGG [21], as well as of complex blocks dedicated to more refined functions; for example, Inception [22,23], which concatenates multiple CONV and POOL blocks to obtain multi-level feature extraction, and ResNet [24], which leverages identity shortcut connections between different blocks to reduce the so-called *vanishing gradient effect* (i.e., the progressive reduction of the gradient error term through a large number of layers). This evolution in depth and complexity is well represented by the chronological list of winners of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [25], a widely-known object recognition contest that has been dominated by CNNs since 2012.

### *3.1. CNN: Full Training*

The most straightforward way of addressing histological classification problems with CNNs is to fully train the network on a set of annotated histological samples (i.e., the colorectal image dataset described in Section 2.1) and then to use it to classify new unlabeled samples, sharing the same characteristics of the ones used for training.

To investigate different depths and architectural characteristics, in our work, we experimented with full training on six different CNN models (see the complete list in Table 2, where SimpleNet is a home-designed shallow network, consisting of just one CONV and one Fully-Connected (FC) layer, and the others are models from the recent literature, with progressively higher depth).

**Table 2.** Full training: CNN models. Each row of the table reports: name, number of Convolutional (CONV) and Fully-Connected (FC) blocks, and, where available, the top-5 % error on the ImageNet dataset, as well as the reference publication.

| | # CONV | # FC | ImageNet Top-5 Error | Ref. |
|---|---|---|---|---|
| SimpleNet | 1 | 1 | n.a. | n.a. |
| LeNet | 2 | 2 | n.a. | [19] |
| AlexNet | 5 | 3 | 16.4% | [20] |
| VGG-16 | 16 | 3 | 7.4% | [21] |
| Inception-v3 | 47 | 1 | 5.6% | [23] |
| ResNet-50 | 49 | 1 | 5.3% | [24] |

[1] home-developed.

All the CNNs were developed within the Keras framework [26], strictly following the implementation described in the corresponding publications (see the last column of Table 2). The networks were trained with a backpropagation paradigm, an iterative process involving multiple passes of the training dataset until the model converges to an optimal configuration of the parameters. At each training step, the whole dataset flows from the first to the last layer in order to compute a classification error, which is quantified by means of a *loss function* (in our implementation, the categorical cross-entropy). Then, the error term flows backward through the net. At each training epoch, the model parameters (i.e., the network weights) are tuned in the direction that

minimizes the classification error on the training data. More specifically, our optimization algorithm applied a Stochastic Gradient Descent (SGD) implemented with a momentum update approach [27] to minimize the categorical cross-entropy function between the five classes of interest. Ten percent of the training set (i.e., *validation set*) was used to monitor the training process and optimize the choice of hyper-parameters of the net. This validation set was completely independent of the images used for testing purposes and was solely used to compute the accuracy metric upon which the training process was optimized. Based on this, we imposed a Learning Rate (LR) equal to 0.0001, a Momentum (M) equal to 0.9, and a Batch Size (BS) of 32 images. To decrease the computational costs and reduce overfitting, we implemented an *early stopping* criterion, which interrupts the training process when validation accuracy does not improve for 10 subsequent epochs [28]. To ensure an efficient exploration of the solution space, we also applied a variable LR strategy, progressively reducing LR each time the validation accuracy did not improve for 5 consecutive epochs.

For the full training approach, CNNs' weights were randomly initialized and then trained on the colorectal cancer training dataset described in Section 2.1. The training was performed on a Linux Infiniband-QDR MIMD Distributed Shared-Memory Cluster provided with a single GPU (NVIDIA Tesla K40, 12 GB, 2880 CUDA cores).

### 3.2. CNN: Transfer Learning

As discussed in the previous sections, a CNN can be seen as a cascade of trainable filter banks, where the successive blocks are devoted to detecting patterns at an increasing level of abstraction, from the lowest (i.e., edges or simple shapes), to the highest (objects and complex shapes). Hence, while the top-most blocks are tailored to a specific classification task, the lower-level ones are ideally generalizable to a large number of applications. Based on this concept, a CNN that was trained on a certain dataset can be transferred to a different context or even used as a feature generator for more than a classification task. This approach, which goes by the name of *transfer learning*, potentially solves the issue of fully training the network on a huge number of labeled training images.

In our work, we chose as the base for transfer learning a VGG-16 architecture, which in our preliminary experiments provided the best trade-off between representation depth, computational costs, and simplicity of interpretation [21]. In spite of its depth, VGG-16 is indeed very simple architecture-wise, as it consists of a linear sequence of convolution and pooling blocks all of the same size and characteristics ($3 \times 3$ and $2 \times 2$, respectively) and a three-layered Fully-Connected block (FC). Conceptually, the locally-connected stages can be represented as a simple sequence of 5 macro-blocks, each ending with a POOL layer (see Figure 4). This linear structure is very convenient for our study, as it allows easy interpretation of the feature extraction process, as well as of the impact of successive blocks on the development of image representation.

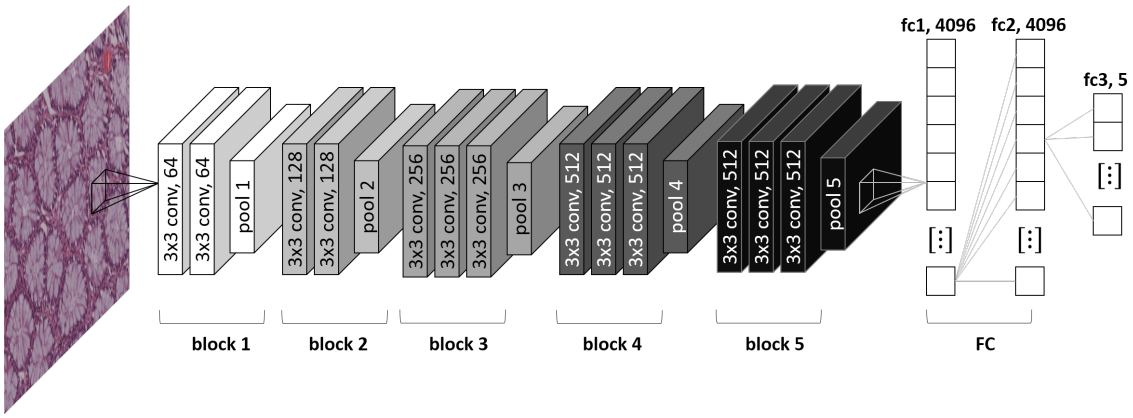


**Figure 4.** Architectural representation of the VGG-16 deep network used for transfer learning (main building blocks).

All the non-linear transforms in the model were ReLU, except for the last fully-connected layer, which had a softmax activation function. The stride was fixed to 1 pixel for convolution and padding and to 2 pixels for max pooling. Differently from the original VGG-16 model, in our implementation, the size of the last fully-connected layer was fixed to 5, matching the number of categories of our classification problem.

In our transfer learning experiments, the VGG-16 network was pre-trained on the ImageNet dataset from the Large-Scale Visual Recognition Challenge 2012 (ILSVRC-2012), which contains 1.2 million photographs depicting 1000 different categories of natural objects [29]. Hence, the domain on which the model was trained completely differed from histopathological image classification, both in terms of image content and characteristics, as well as in terms of the number of classes.

To investigate the possibility of transferring the model from ImageNet to our classification task fully, we designed and compared three different approaches (see Figure 5):

1. *CNN as a fixed feature generator.* The histological images were fed into the pre-trained CNN only for inference. The features extracted by the convolutional blocks were then fed into a separate machine learning framework, consisting of a feature reduction stage and a supervised classifier.
2. *Partial fine-tuning of the CNN.* The CNN model learned on the ImageNet was re-trained on our training set of histological images, keeping all the parameters of a few low-level blocks fixed to their initial value. Hence, only the weights of a certain number of top-most layers were fine-tuned for histological image classification.
3. *Complete fine-tuning of the CNN.* As for Approach (b), the CNN model was initialized with the values learned on the ImageNet. Then, all the blocks of the CNN (including the low-level ones) were re-trained on the histological images.

#### 3.2.1. CNN as a Fixed Feature Generator

As the first transfer learning methodology, the CNN with parameters learned on the ImageNet was used as-is to infer the image descriptors for the new classification problem. For this purpose, we used the output of the POOL5 layer of the pre-trained CNN as features for histological image classification (details will follow). The feature vector was fed into a separate machine learning framework, as represented by Figure 5a. This framework consisted of the following steps:

1. *Feature reduction.* To reduce the dimensionality of the data and prevent overfitting, we applied a Principal Component Analysis (PCA). PCA is a well-established method that orthogonally transforms the original features into a new group of values, which are linear combinations of the original characteristics, the so-called principal components. As the transformation works towards minimizing the correlation between the features, the new data representation is expected to best summarize those features that are most representative of the classes of interest.
2. *Classification.* The final classification into five categories (H, T, V, S, AC) was performed by a Support Vector Machine (SVM) with a Gaussian radial basis function kernel. The hyper-parameters of the kernel were set by means of a Bayesian Optimization (BO) algorithm [30], implementing a 10-fold cross-validation procedure on the training images. This procedure was found to provide much better and faster results compared to classic methods based on grid search or heuristic techniques.

The parameters of the framework were empirically established on a subset of the training images, as follows. First, we ran experiments varying the CNN block used as the feature generator (from POOL1 to FC2, respectively) and quantified the accuracy of the SVM at an increasing number of principal components imposed on the PCA. The results of this experiment are in Figure 6a. In this graph, the per-class accuracies are reported by means of bars that extend from the minimum to the maximum value obtained by each of the five categories (H, T, V, S, and AC) at different numbers of principal components (100, 500, and 900, respectively). As can be observed, extracting features from POOL5 and reducing the number of features to 100 principal components ensured the best accuracy among

all the classes (i.e., the lower-end of the bar was the highest). To further refine the optimization, the experiment was repeated with only features from the POOL5 layer, but with a much finer resolution of the number of principal components. By doing so, we obtained that the optimal number of principal components was 200 (see Figure 6b).

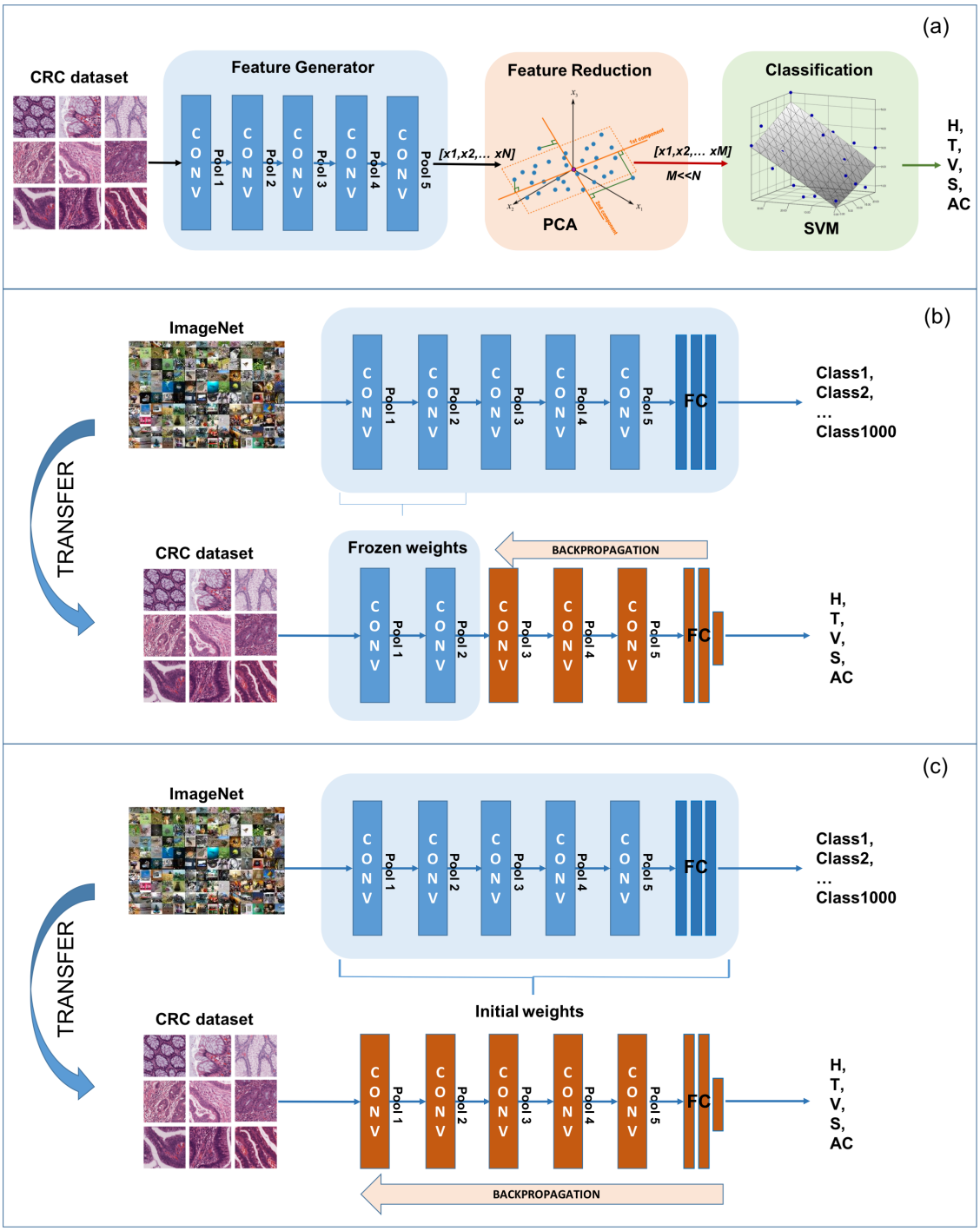


**Figure 5.** Schematic representation of the three transfer learning techniques designed and compared in this work: (**a**) Pre-trained CNN as a fixed feature generator. (**b**) Partial fine-tuning of pre-trained CNN. (**c**) Complete fine tuning of pre-trained CNN. CRC, Colorectal Cancer; H, Hyper-plastic polyp; T, Tubular adenoma; V, Villous adenoma, S, Serrated adenoma; AC, Adenocarcinoma.

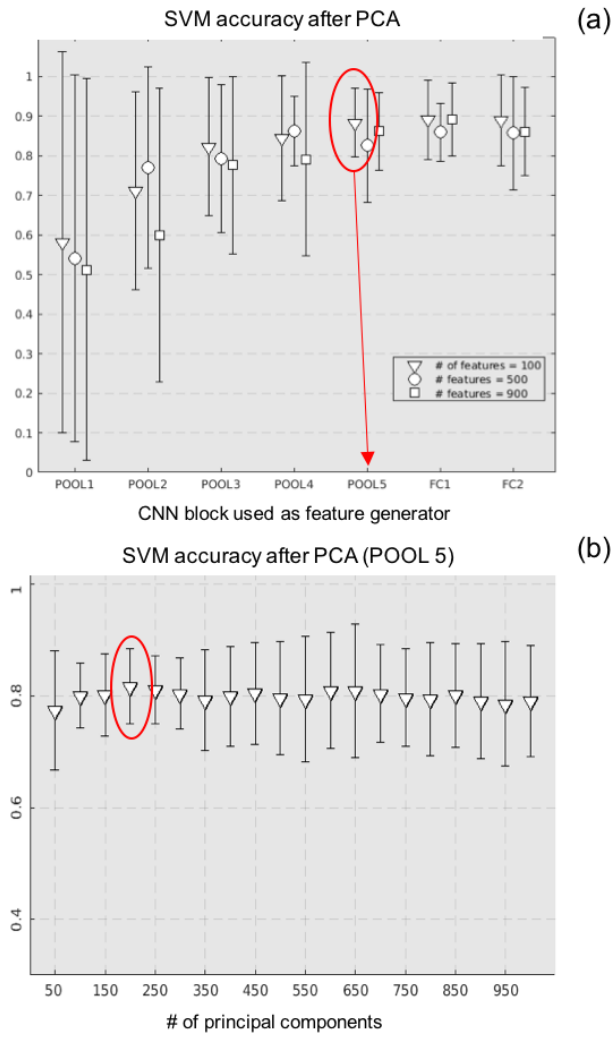


**Figure 6.** Optimization of the classification framework. (**a**) SVM accuracy versus the CNN block used as the feature generator, at different values of principal components imposed on the PCA. The bars extend from the minimum to the maximum accuracy value obtained by each of the five polyp categories (best configuration circled in red). (**b**) SVM accuracy versus the number of principal components imposed on the PCA, with the POOL5 block used as the feature generator.

### 3.2.2. Partial Fine-Tuning of Pre-Trained CNN

As a second transfer learning methodology, we tried adapting the pre-trained VGG-16 to our specific classification task. For this purpose, we first initialized all the weights of the network to the ones determined on the ImageNet dataset, as represented in Figure 5b. Then, we continued the backpropagation procedure on our histological dataset, keeping the weights of the first blocks of the net (more specifically, POOL1 and POOL2) frozen. The rationale of such a strategy is trying to maintain the low-level features describing the most generic and generalizable details (e.g., edges and simple shapes) as they were learned from the ImageNet. Hence, all the computational power can be devoted to the training of the top-most layers, which are expected to learn high-level task-specific features. The training strategy was exactly the same as was described in Section 3.1 and lasted 4 h on the same hardware. As for the previous transfer learning strategy, the starting block for the backpropagation was decided by running experiments on a subset of the training images. As can be seen from Figure 7, POOL2 was the block ensuring the best accuracy across the five categories.

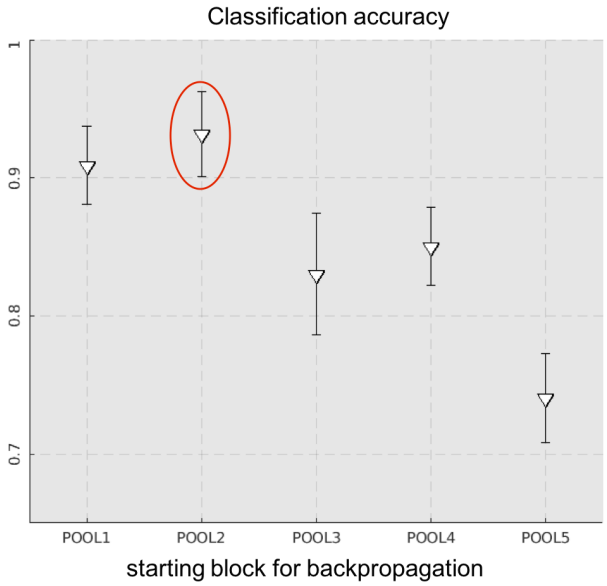


**Figure 7.** Optimization of the partial fine-tuning. Classification accuracy versus the starting block of the backpropagation (the bars extend from the minimum to the maximum accuracy value obtained by each of the five polyp categories).

#### 3.2.3. Complete Fine-Tuning of Pre-Trained CNN

As a third transfer learning methodology, we extended the fine-tuning to all the blocks of the pre-trained VGG-16, using the weights learned on the ImageNet just for initialization (see Figure 5c). Again, the training was performed by backpropagation and lasted 5.5 h.

### *3.3. Traditional Machine Learning Approach*

In order to provide a benchmark to our CNN-based classifiers, we designed and implemented a traditional machine learning approach based on handcrafted feature extraction. More specifically, we implemented a *Bag Of Features* (BOF) framework leveraging SURF, a scale- and rotation-invariant keypoint detector and descriptor [6,31], and a support vector machine classifier. This is a consolidated approach to histological image classification [4,5].

In our BOF framework, the local SURF descriptors were first extracted from the training images and then grouped into clusters by means of a k-means clustering algorithm. The clusters' centroids were used to generate a codebook of so-called *visual words*, upon which image representation can be built (see Figure 8). The histogram of occurrences of the visual words was then used as the feature vector for the classification, which was performed by a classic SVM classifier optimized via the same procedure described in Section 3.2.1.

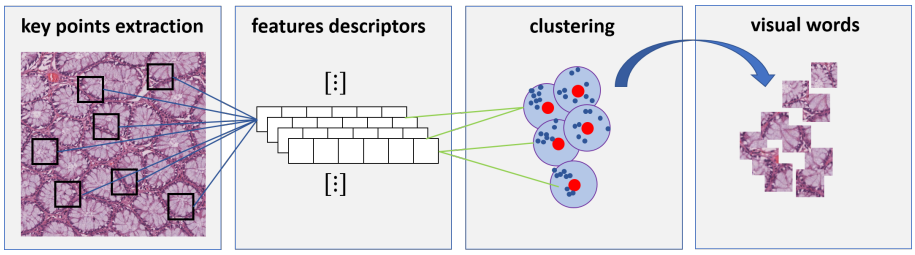


**Figure 8.** Bag of features approach: image representation.

## 4. Results

Classification performance was established using the test dataset of colorectal samples described in Section 2.1. As already mentioned, this set was completely independent of the one used for optimizing

and training the classifiers (i.e., the patches were obtained from different patients). As for the training set, the test set was balanced, in that it contained an equal number of patches of the five different polyp categories.

The classification accuracy was quantified both in terms of patches and of patients correctly classified, following the same notation of [32]. More specifically, we extracted two different accuracy metrics, namely *patch score* ($S_P$) and *patient score* ($S_{Pt}$).

The *patch score* ($S_P$) is a measure of patch-wise classification accuracy, defined as the fraction of patches of the test set that were correctly classified:

$$S_P = \frac{N_C}{N}, \tag{1}$$

where $N_C$ is the number of test patches correctly classified and $N$ the total number of patches in the test set.

The *patient score* ($S_{Pt}$) is related to the patient-wise accuracy and is a derivation of the patch score. It is obtained by computing the fraction of patches belonging to a single patient (i.e., to the same WSI) that were correctly classified. This value, which can be defined as a *per-patient* patch score, is then averaged over all the patients in the test set, as follows:

$$S_{Pt} = \frac{\sum_i S_P(i)}{N_P}, \tag{2}$$

where $S_P(i)$ is the patch score of the $i^{\text{th}}$ patient and $N_P$ the total number of patients in the test set.

In Table 3, we report both the patch and patient scores (mean ± standard deviation) obtained in our experiments. More specifically, the first and the second part of the table show, respectively, the accuracy values of the CNNs fully trained on colorectal histological images (Section 3.1) and of the CNNs trained with transfer learning methodologies (Section 3.2), with the latter reported as follows:

1. *CNN + SVM* refers to the SVM classifier, with the pre-trained CNN used as fixed feature generator.
2. *fine-tune-CNN (partial)* refers to the pre-trained CNN with only the weights of the top-most layers fine-tuned.
3. *fine-tune-CNN (complete)* refers to the model fully retrained on the CRC training set, with weights initialized based on ImageNet.

Finally, the last row of the table shows the values of the traditional Bag of Features (*BOF + SVM*) framework, taken as a reference for the accuracy assessment on the same test set.

**Table 3.** Classification accuracy: patch and patient scores (mean ± std). BOF, Bag Of Features.

| | | $S_P$ | $S_{Pt}$ |
|---|---|---|---|
| **CNN full training** | *SimpleNet* | 0.60 | 0.63 (± 0.14) |
| | *LeNet* | 0.74 | 0.79 (± 0.28) |
| | *AlexNet* | 0.74 | 0.74 (± 0.30) |
| | *VGG-16* | 0.69 | 0.76 (± 0.37) |
| | *Inception v3* | 0.67 | 0.70 (± 0.40) |
| | *ResNet50* | 0.67 | 0.71 (± 0.40) |
| **CNN transfer learning** | *CNN + SVM* | 0.93 | 0.95 (± 0.07) |
| | *fine-tune-CNN (partial)* | 0.93 | 0.93 (± 0.01) |
| | *fine-tune-CNN (complete)* | 0.96 | 0.96 (± 0.08) |
| **Traditional ML** | *BOF + SVM* | 0.83 | 0.82 (± 0.14) |

Quite interestingly, irrespective of the depth and architectural complexity, none of the CNNs fully trained on the colorectal histological dataset were able to match the accuracy of the BOF + SVM framework. The accuracy of the fully-trained CNNs was at least 10% lower than the traditional ML

technique (which obtained a 83% patch score), with very high variability of accuracy from patient to patient (standard deviation of patient score spanning from 14–40%). These values suggest that the CNNs were not able to build a generalizable image representation on the given training set, most probably due to the high variability of the image characteristics and the relatively low number of patients used for training. As a matter of fact, even the BOF + SVM technique, which is generally less susceptible to small training sets than CNNs, obtained remarkable variability of the outcome from patient to patient (14% patient score standard deviation).

The outcome of the transfer learning techniques, on the other hand, was surprisingly good. Both the patch and patient scores were consistently high (above 93%) for all the tested methodologies, overcoming the BOF + SVM method by 10% at least. Furthermore, the accuracy computed over all the patches of the test set was very similar to the one computed patient per patient, with a much smaller standard deviation of the latter value. This suggests that all the transfer learning classification frameworks, unlike CNNs trained from scratch, were reasonably robust and coped well with inter-patient variability.

The same considerations hold when analyzing the outcome of the experiments on a class-per-class basis. Figure 9 shows the patch-wise classification performance of all the classification models reported in Table 3, in the form of $5 \times 5$ confusion matrices. Each row of a matrix represents the fraction of patches in a predicted class (respectively, AC, H, S, T, or V), while each column represents the fraction of instances in a true class. Hence, the main diagonal of the matrix collects the correct classifications (i.e., the instances where the predicted class coincided with the actual class), while the rest of the elements in the matrix are classification errors.

Again, the matrices show that the fully-trained CNNs (see (a) of the figure) obtained inconsistent classification results on different classes of polyps. Serrated adenomas, which are generally reported in the literature to be difficult to identify [33], were misclassified by all the CNNs irrespective of their depth. Even more remarkably, the adenocarcinoma class (most probably the one with the highest variability in terms of morphological and architectural characteristics of the tissue) also obtained low classification accuracy compared to the other benign categories of polyps. Indeed, this nullifies the practical diagnostic usability of the classification frameworks based on CNNs trained from scratch. On the other hand, the BOF + SVM methodology ((c) of the figure) obtained much better and more homogeneous results on the five different classes than the CNNs trained from scratch. Still, the accuracy on adenocarcinoma was quite low (60%).

The high accuracy of the transfer learning methodologies is again confirmed by the class-per-class analysis. As can be easily gathered from the matrices in Figure 9b, the accuracy values were consistently high for all five polyp categories. Among the transfer learning methods, the SVM-based classifier with CNN used as the feature extractor had slightly lower accuracy than the others on serrated adenomas, but still close to 80%. Both the fine-tuning methodologies (partial and complete) obtained accuracy higher than 80% (and most of the times higher than 90%) for all the polyp categories. Interestingly, the worst and best performing class was not always the same: for example, the hyperplastic polyp obtained the lowest accuracy value (83%) with the partial fine-tuning and the highest accuracy (100%) with the complete fine-tuning. Fully retraining the CNN model using the weights learned from ImageNet for initialization obtained the best results in terms of accuracy and consistency of performance among the five classes of polyps. By comparing the confusion matrices of (b) with the one of the BOF + SVM methodology, we can observe that the description capability of the features obtained from transfer learning was better than the traditional multi-purpose image descriptors, even though the transfer was from a completely different imaging domain.

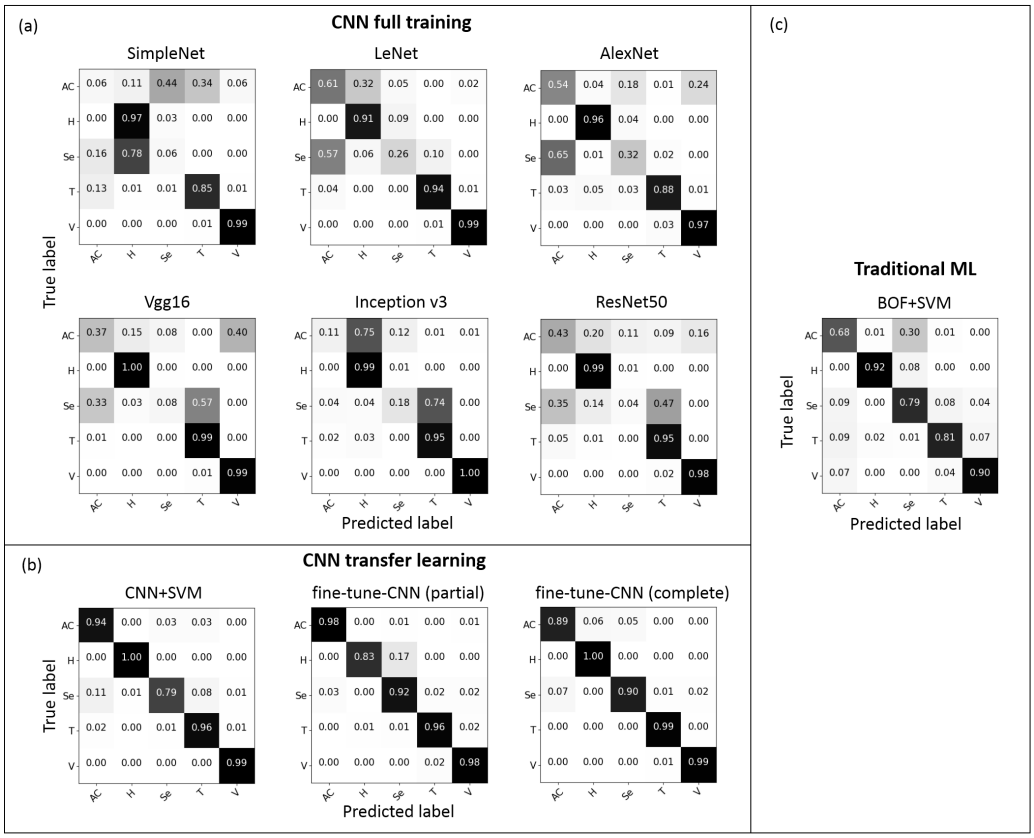


**Figure 9.** Patch-wise confusion matrices (%) of all the classification frameworks. (**a**) CNNs fully trained on histopathological images. (**b**) Transfer learning methodologies. (**c**) Classification framework based on bag of features and support vector machine.

*Additional Experiments on Fine-Tuning*

Based on our experiments on the colorectal dataset, in the case of the unavailability of large training datasets, CNNs pre-trained on a completely different imaging context (i.e., the ImageNet) are a good solution to the problem of histological image classification. More specifically, a pre-trained VGG-16 network, after a complete or even partial fine-tuning on the histological dataset, is a good alternative to a traditional feature design approach (e.g., the BOF + SVM model).

In order to prove the generality of these findings, we experimented with the same fine-tuning approach (i.e., VGG-16 architecture, pre-trained on the ImageNet) on two additional histological datasets, respectively from cardiovascular and bone tissues (see Figure 3a,b). As already discussed in Section 2.2, these two datasets are the ones that are most different from our main case study (i.e., colorectal tissue), as well as from each other.

As the two datasets are available only in the form of patches and not of full WSIs, we evaluated classification accuracy only in terms of patch scores (Equation (1)), reporting the obtained values in Figure 10. More specifically, the graph shows the patch-score obtained by the pre-trained CNN with different configurations of the fine-tuning (that is, with re-training of the complete network or of progressively smaller portions of the network). As for the other graphs in Section 3, the $X$ axis shows the starting block of the backpropagation algorithm (that is, all the preceding blocks are fixed to their initial values learned on the ImageNet throughout the learning process), and the $Y$ axis shows the corresponding patch score. The plotted lines have markers corresponding to the average accuracy value and bars spanning from the minimum to the maximum of the accuracy values on each individual class.

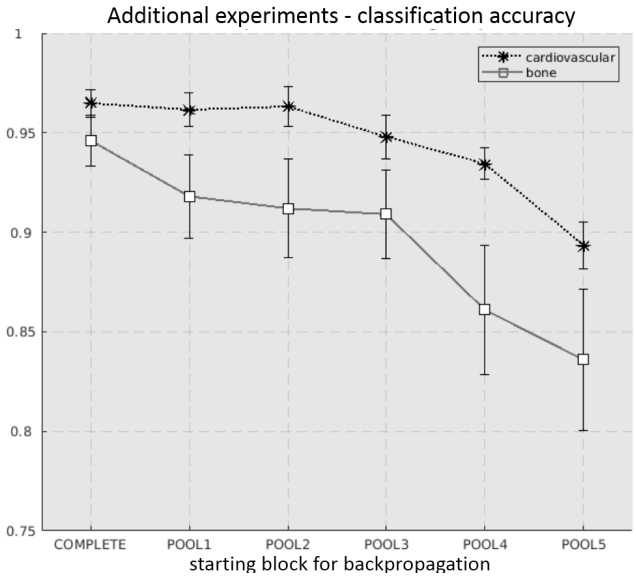


**Figure 10.** Patch scores of a fine-tuned VGG-16 network on the *cardiovascular* and *bone* datasets, at different extensions of the fine-tuning. The markers show the average accuracy values, while the bars extend from the minimum to the maximum accuracy values on each individual class.

For both the cardiovascular and the bone tissue datasets, the trend of the accuracy values was not very different from the one obtained on the colorectal images: the accuracy tended to decrease when the backpropagation was applied to progressively smaller portions of the network. This is not surprising, as the fine-tuning allows creating more problem-specific features than the ones learned on ImageNet. On the other hand, the accuracy drop was fairly small when the first three blocks of the networks were left unchanged. Hence, partial fine-tuning can be applied in the case of a lack of computational time and training resources, without significant loss in terms of classification performance. As can be gathered from the graph, transfer learning obtained very good classification results (the complete fine-tuning obtained 97% and 94%, respectively, for the cardiovascular and the bone dataset). According to our tests, the BOF + SVM approach had on average accuracy values 15% lower than the fine-tuned CNNs.

The same considerations can be drawn analyzing the per-class accuracies. Figure 11 shows confusion matrices of the CNNs with complete fine-tuning (and, for comparison, of the BOF + SVM methodology), respectively, for the cardiovascular and for the bone tissue dataset. Once again, the fine-tuned CNNs obtained higher classification performance and higher consistency of the accuracy on different histological classes.

Based on our results, complete fine-tuning is the transfer learning strategy leading to the best classification accuracy. However, how does fine-tuning from a completely different domain (e.g., ImageNet) compare with fine-tuning from a similar image domain, say histopathological images from a different tissue? To answer such a question, we used as a case study the pleural tissue dataset described in Section 2.2 (see Figure 3c), always using VGG-16 as the CNN architecture. As they both contain epithelial tissue samples, the pleural and the colorectal tissue datasets are in theory the most similar from a histological point of view. Hence, we investigated the possibility of transferring the CNN across these two histological classification problems.

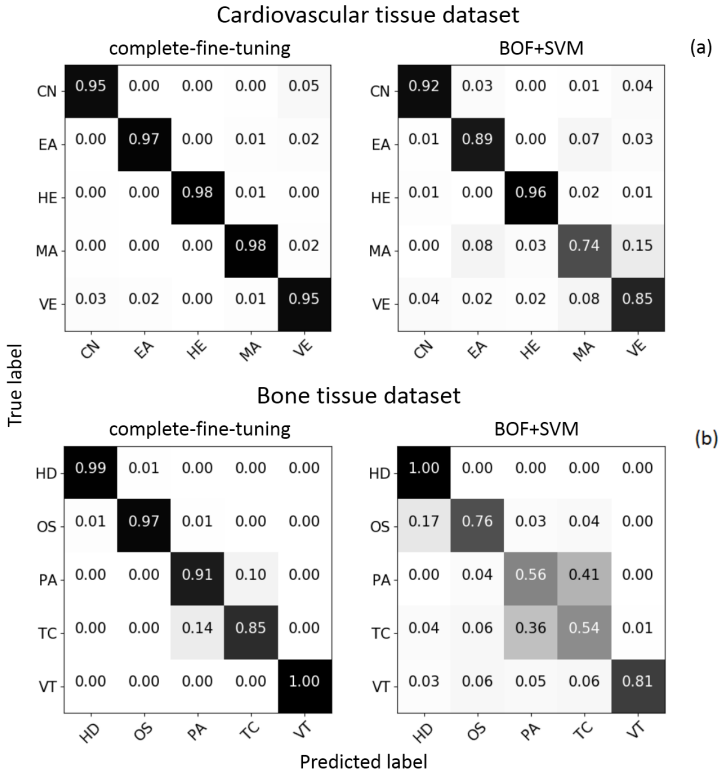


**Figure 11.** Confusion matrices on the *cardiovascular* (**a**) and *bone* (**b**) tissue datasets, respectively of a CNN with complete fine-tuning and of a BOF + SVM framework.

In the columns of Table 4, we report the obtained pleural tissue classification accuracy values, respectively:

1. fully training the CNN on the pleural tissue images;
2. initializing the CNN with weights learned on the ImageNet (*ImageNet fine-tuning*) and then fine-tuning such weights on the pleural images;
3. initializing the same CNN with the weights that obtained the best classification accuracy on the CRC dataset, as reported in Section 4 (*CRC fine-tuning*).

Even in this case, we report the total accuracy value $\pm$ the span of the accuracy values on each individual class of the pleural dataset.

**Table 4.** Pleural tissue classification accuracy (%). Comparing cross-domain and intra-domain transfer learning strategies.

| Full Training | ImageNet Fine-Tuning | CRC Fine-Tuning |
|---|---|---|
| $75 \pm 1\%$ | $80 \pm 2\%$ | $77 \pm 2\%$ |

As can be gathered from the table, both the transfer learning methods overcame full training, which is consistent with our previous results. On top of that, fine-tuning a network pre-trained on a completely different domain (ImageNet) obtained even better results than fine-tuning a network pre-trained on a different histological dataset (80% against 77%). This further confirms that, provided that the original training domain is sufficiently large, miscellaneous, and general in terms of image characteristics, cross-domain transfer learning is a feasible solution to the problem of histological image classification with CNNs.

## 5. Discussion

Analyzing the outcome of our experiments, we can draw the following considerations.

The results obtained training the CNNs from scratch were rather disappointing, regardless of the depth and of the architectural complexity of the model. Indeed, none of the tested architectures ensured accurate classification of all the categories of interest. A reasonable explanation of this result is the paucity of the training set, not in terms of the number of image patches (which is comparable to most literature approaches), but in terms of the number of patients per class. Nonetheless, obtaining a much larger training dataset is not always a viable solution in a clinical setting, as it may have prohibitive costs both in terms of annotation efforts, as well as computational resources. On top of that, as already discussed, biological tissues have tremendous variability even within the same histological class, and usually, they do not have canonical orientations and shapes. Hence, synthetic data augmentation techniques commonly applied in computer vision tasks (e.g., image rotations, shifting, flipping, etc.) come with a serious risk of over-fitting.

Extracting knowledge from a CNN that has been trained using a large labeled dataset from a different application (i.e., the ImageNet, which contains photographs of every-day objects and natural scenes, and not histological samples) seems a viable alternative to the classic training from scratch or to artificial data augmentation. This contradicts the assumptions of the initial works on transfer learning, according to which the transfer had to involve two similar imaging domains to be successful. The observed results show that the low-level features learned by the first stages of a CNN can be successfully generalized to the context of histological image classification and provide comparable or even better classification performance than traditional handcrafted descriptors. Indeed, using the CNN learned on the ImageNet as a fixed features' generator provided acceptable classification even with a rather simple classifier based on support vector machines. Hence, this may be the preferred solution when the computational burden of the learning procedure and the availability of training data are a serious constraint.

In terms of sheer accuracy, fine-tuning the pre-trained CNN is the transfer learning approach that obtained the best performance, when compared with the traditional BOF + SVM approach, and the complete fine-tuning slightly outperformed the partial one. Indeed, fine-tuning allows extracting more context-specific features than the CNN + SVM approach. On the other hand, it is more time consuming, as it still requires running the backpropagation algorithm on the training set. Nonetheless, the computational time is more than halved compared to training the CNN from scratch and ensures good classification with relatively smaller training sets.

In general, fine-tuning works best when it is applied to the entire deep network. Nevertheless, our experiments on multiple histological datasets showed that the backpropagation can be limited to a reduced number of high-level blocks, without impacting the classification performance too much. The extension of the fine-tuning can be set with a cost-benefit analysis, based on the extent of the training set and the computational resources available.

## 6. Conclusions

The purpose of the current study was to investigate the practical use of deep learning, and more specifically of convolutional neural networks, for the automatic classification of histopathological images, which is a task characterized by very high intra-class variability. To this aim, we tested the performance of several CNN models on the task of colorectal polyp assessment, which is a very challenging and important multi-class classification problem in histopathology. CNNs are, in theory, ideal for this classification task, as they avoid the extraction of a fixed set of handcrafted features. Nonetheless, our experiments revealed a non-satisfactory performance of the CNNs trained from scratch, regardless of the depth and architectural complexity of the model. The limited number of training examples, coupled with the high complexity and variability of the image characteristics, was held responsible for the bad performance of the deep neural network approach.

As an alternative to training the CNN from scratch on the histological samples, we investigated the possibility of using transfer learning techniques, based on a CNN model pre-trained on a completely different classification problem (i.e., the ImageNet). More specifically, we designed and experimentally compared three different techniques involving transfer learning, namely (i) using the pre-trained CNN to extract features for a separate machine learning framework, consisting of a PCA for dimensionality reduction and a SVM as classifier, (ii) fine-tuning on histopathological images a selected number of blocks of the pre-trained CNN, and (iii) fine-tuning the whole pre-trained CNN, using a priori knowledge from ImageNet just for weights' initialization.

In our experiments, all the transfer learning methodologies outperformed the CNN trained from scratch, as well as a traditional BOF + SVM framework, showing that using CNN models learned on a completely different dataset and classification context is a viable solution and solves the issues of having very large annotated datasets for the training. Among the transfer learning methodologies, the complete fine-tuning was the one obtaining the best results in terms of mean and per-class accuracy. On the other hand, even the other transfer learning techniques obtained reasonably good classification performance, with even lesser computational time. The same conclusions were confirmed by additional experiments on different histological problems.

In conclusion, our findings validate the potentials of transfer learning methodologies for the automated classification of histopathological images, leveraging models learned on a completely different classification problem. This solves the problem of the unavailability of large annotated training sets, as well as of computational resources for the training and finally opens the way toward the practical and efficient exploitation of convolutional neural networks in computer-aided diagnosis systems for digital pathology.

**Author Contributions:** Conceptualization, S.D.C. and E.F.; methodology, S.D.C. and F.P.; software, F.P.; validation, F.P.; writing, original draft preparation, S.D.C., F.P., and G.U.; writing, review and editing, S.D.C., G.U., and E.F.; supervision, S.D.C. and E.F.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. He, L.; Long, L.R.; Antani, S.; Thoma, G.R. Histology Image Analysis for Carcinoma Detection and Grading. *Comput. Methods Prog. Biomed.* **2012**, *107*, 538–556. [CrossRef] [PubMed]
2. Farris, A.B.; Cohen, C.; Rogers, T.E.; Smith, G.H. Whole Slide Imaging for Analytical Anatomic Pathology and Telepathology: Practical Applications Today, Promises, and Perils. *Arch. Pathol. Lab. Med.* **2017**, *141*, 542–550. [CrossRef] [PubMed]
3. Young, A.; Hobbs, R.; Kerr, D. *ABC of Colorectal Cancer*, 2nd ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2011.
4. Komura, D.; Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Comp. Struct. Biotechnol. J.* **2018**, *16*, 34–42. [CrossRef] [PubMed]
5. Di Cataldo, S.; Ficarra, E. Mining textural knowledge in biological images: Applications, methods and trends. *Comp. Struct. Biotechnol. J.* **2017**, *15*, 56–67. [CrossRef] [PubMed]
6. Panchal, P.; Panchal, S.; Shah, S. A comparison of SIFT and SURF. *Int. J. Innovat. Res. Comp. Commun. Eng.* **2013**, *1*, 323–327.
7. Kather, J.; Weis, C.A.; Bianconi, F.; Melchers, S.; Schad, L.; Gaiser, T.; Marx, A.; Zöllner, F. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* **2016**, *6*, 27988. [CrossRef] [PubMed]
8. Csurka, G.; Dance, C.R.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, 10–16 May 2004; pp. 1–22.
9. Caicedo, J.C.; Cruz, A.; Gonzalez, F.A. Histopathology Image Classification Using Bag of Features and Kernel Functions. *Artif. Intel. Med.* **2009**, 126–135. [CrossRef]

10. Janowczyk, A.; Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **2016**, *7*, 29. [CrossRef] [PubMed]
11. Korbar, B.; Olofson, A.M.; Miraflor, A.P.; Nicka, C.M.; Suriawinata, M.A.; Torresani, L.; Suriawinata, A.A.; Hassanpour, S. Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. *J. Pathol. Inform.* **2017**, *8*, 30. [CrossRef] [PubMed]
12. Vununu, C.; Lee, S.H.; Kwon, K.R. A Deep Feature Extraction Method for HEp-2 Cell Image Classification. *Electronics* **2019**, *8*, 20. [CrossRef]
13. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
14. Ribeiro, E.; Uhl, A.; Wimmer, G.; Hafner, M. Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification. *Comp. Math. Method. Med.* **2016**, *2016*, 1–16. [CrossRef] [PubMed]
15. Xu, Y.; Jia, Z.; Wang, L.B.; Ai, Y.; Zhang, F.; Lai, M.; Chang, E.I.C. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinform.* **2017**, *18*, 281. [CrossRef] [PubMed]
16. Ponzio, F.; Macii, E.; Ficarra, E.; Di Cataldo, S. Colorectal Cancer Classification using Deep Convolutional Networks—An Experimental Study. In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies–Volume 2: BIOIMAGING, INSTICC, SciTePress, Madeira, Portugal, 19–21 January 2018; pp. 58–66. [CrossRef]
17. Haggar, F.A.; Boushey, R.P. Colorectal Cancer Epidemiology: Incidence, Mortality, Survival, and Risk Factors. *Clin. Colon Rectal Surg.* **2009**, *22*, 191–197. [CrossRef] [PubMed]
18. Mazo, C.; Alegre, E.; Trujillo, M. Classification of cardiovascular tissues using LBP based descriptors and a cascade SVM. *Comp. Method. Prog. Biomed.* **2017**, *147*, 1–10. [CrossRef] [PubMed]
19. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, 1097–1105. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–12 June 2015; pp. 1–9. [CrossRef]
23. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NE, USA, 26 June–1 July 2016; pp. 2818–2826.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NE, USA, 26 June–1 July 2016; pp. 770–778.
25. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comp. Vision (IJCV)* **2015**, *115*, 211–252. [CrossRef]
26. Chollet, F. Keras, GitHub. 2015. Available online: https://github.com/fchollet/keras (accessed on 1 January 2019).
27. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Networks* **1999**, *12*, 145–151. [CrossRef]
28. Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Construc. Approx.* **2007**, *26*, 289–315. [CrossRef]
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR09, Miami, FL, USA, 20–25 June 2009.
30. Hastie, T.; Tibshirani, R.; Friedman, J. Overview of Supervised Learning. In *The Elements of Statistical Learning*; Springer: Berlin, Germany, 2009.
31. Juan, L.; Gwun, O. A comparison of sift, pca-sift and surf. *Int. J. Image Process. (IJIP)* **2009**, *3*, 143–152.

32. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. Breast cancer histopathological image classification using convolutional neural networks. In Proceedings of the 2016 IEEE International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2560–2567.
33. Mohamed, M.; Schofield, J.B. The pathology of colorectal polyps and cancers (including biopsy). *Surgery* **2014**, *32*, 165–171. [CrossRef]