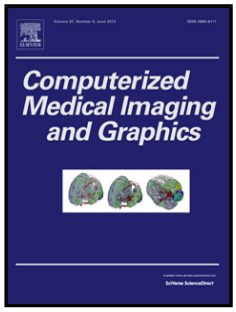


Title: Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology

Author: Harshita Sharma Norman Zerbe Iris Klempert Olaf Hellwich Peter Hufnagl

PII: S0895-6111(17)30050-2
DOI: http://dx.doi.org/doi:10.1016/j.compmedimag.2017.06.001
Reference: CMIG 1514

To appear in: *Computerized Medical Imaging and Graphics*

Received date: 8-7-2016
Revised date: 26-4-2017
Accepted date: 8-6-2017

Please cite this article as: Harshita Sharma, Norman Zerbe, Iris Klempert, Olaf Hellwich, Peter Hufnagl, Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology, *<![CDATA[Computerized Medical Imaging and Graphics]]>* (2017), http://dx.doi.org/10.1016/j.compmedimag.2017.06.001

Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology

Harshita Sharma[a,*], Norman Zerbe[b], Iris Klempert[b], Olaf Hellwich[a], Peter Hufnagl[b]

[a] *Computer Vision and Remote Sensing, Technical University Berlin, Berlin, Germany*
[b] *Department of Digital Pathology and IT, Institute of Pathology, Charité University Hospital, Berlin, Germany*

Abstract

Deep learning using convolutional neural networks is an actively emerging field in histological image analysis. This study explores deep learning methods for computer-aided classification in H&E stained histopathological whole slide images of gastric carcinoma. An introductory convolutional neural network architecture is proposed for two computerized applications, namely, *cancer classification* based on immunohistochemical response and *necrosis detection* based on the existence of tumor necrosis in the tissue. Classification performance of the developed deep learning approach is quantitatively compared with traditional image analysis methods in digital histopathology requiring prior computation of handcrafted features, such as statistical measures using gray level co-occurrence matrix, Gabor filter-bank responses, LBP histograms, gray histograms, HSV histograms and RGB histograms, followed by random forest machine learning. Additionally, the widely known AlexNet deep convolutional framework is comparatively analyzed for the corresponding classification problems. The proposed convolutional neural network architecture reports favorable results, with an overall classification accuracy of 0.6990 for cancer classification and 0.8144 for necrosis detection.

Keywords: Deep learning, Convolutional neural networks, Gastric carcinoma, Digital pathology, Histopathological image analysis, Cancer classification, Necrosis detection

*Corresponding author
Email address: harshita.sharma@campus.tu-berlin.de (Harshita Sharma)
URL: www.cv.tu-berlin.de (Harshita Sharma)

2016 MSC: 00-01 99-00

1. Introduction

Gastric cancer is a major cause of cancer and cancer-associated deaths in the world [1]. Computer-based analysis of gastric cancer tissue images is a growing area of research in digital histopathology. Pathologists in routine practice visually navigate and inspect glass slides or whole slide images (WSI) to identify and analyze abnormalities, which is a prolonged and tedious process. Moreover, human eye is less adept to recognize subtle changes in the tissue that may lead to different interpretations among medical professionals, hence, may introduce inter-and intra-observer variability. In this paper, the authors attempt to overcome such problems by proposing a computer-based method using deep learning for histological image analysis in gastric cancer WSI. Cancer classification can potentially assist pathologists in computer-aided diagnosis in the more routinely used H&E stain without the requirement of immunohistochemical staining, thereby reducing costs and efforts in preparation and inspection. Automatic necrosis detection can also decrease viewing times and play an important role in diagnosis and prognosis. The methods can collectively contribute towards reduction in observer variabilities and can be extended to other histological datasets.

Recent literature consists of related works illustrating techniques for application -specific analysis of gastric tissue images in digital histopathology. For instance, image regions depicting normal mucosa, gastritis and adenocarcinoma in H&E stained histological sections are distinguished by use of cytometric measurements in [2]. Gastric atrophy is quantitatively analyzed in H&E stained sections using syntactic structure methods in [3]. A semi-supervised approach for detection and diagnosis of gastric cancer is described in [4] using multiple instance learning in H&E stained tissue images. A multi-resolution method to improve cell nuclei segmentation of gastric cancer is given in [5]. Our work in [6] involves graph-based analysis of H&E stained gastric carcinoma WSI based on their HER2 immunohistochemistry (IHC) and AdaBoost classification. We have also explored necrosis detection in gastric carcinoma using texture features and SVM-based classification in [7]. In this paper, we extend our findings of [6] and [7] by exploring deep learning approaches and quantitatively comparing their performance with the traditional image analysis methods in digital histopathology.

Deep learning using convolutional neural networks (CNN) has lately drawn substantial attention of the scientific community in diverse fields of image analysis [8]. Earliest applications include general object categorization [9] and handwritten zip code classification [10]. Recent advances based on deep learning have even gained prominence in the field of digital histopathology. For example, a CNN method is proposed in [11] for classification of invasive ductal carcinoma in breast cancer WSI with results superior to handcrafted features. Another work using deep neural networks [12] explores detection of mitosis in breast cancer images. U-nets have recently become popular for biomedical image segmentation [13], as these were initially used for segmenting neuronal structures in electron microscopy stacks and further applied to transmitted light microscopy images to perform cell tracking. A comparative study of deep CNN architectures with handcrafted methods for classification of stromal and epithelial histological images of breast cancer and colorectal cancer is given in [14]. A patch-based convolutional neural network approach is examined in [15], for discriminating glioma and non-small-cell lung carcinoma into respective subtypes in histological WSI. However, deep learning using convolutional neural networks in digital histopathology is still in its early stages of development, and this work contributes towards this direction for image analysis in gastric cancer WSI.

There are several motivations to perform this study. Firstly, in the knowledge of the authors, deep learning methods have not been explored until now for the specified classification problems, namely, cancer classification based on IHC and necrosis detection in H&E stained histopathological WSI of gastric cancer. Furthermore, one of the most significant promises of deep convolutional neural networks is to replace the requirement of handcrafted feature design and extraction with efficient task-specific learning algorithms. Such algorithms have even been shown to outperform many hand-engineered features in several fields [16], including a few studies in histopathological image analysis. Hence, the authors are highly motivated to explore the potential of deep learning methods by using a basic self-designed CNN architecture for gastric cancer image analysis, and comparing its performance with common handcrafted features, and AlexNet CNN framework *i.e.* extremely successful in general object categorization. The authors also want to examine the generalizability of deep learning methods by evaluating the same learning machines for two histological image analysis tasks, namely, cancer classification and necrosis detection. Additionally, deep learning studies often use

smaller images for training purpose, however, in our experiments, high resolutions (512× 512 pixels) provide a larger field of view that is preferable for analysis of histological images, as it helps acquiring context information such as neighborhood properties and tissue architecture at higher magnification (40×), but is not widely explored due to memory restrictions. Lastly, training and classification based on deep learning includes direct processing of image regions and eliminates the requirement of a segmentation stage, hence, its performance is not limited by the results of a cell nuclei segmentation algorithm, as observed in [6].

The outline of the paper is as follows. Section 2 describes the background of deep learning and traditional machine learning methods explored in this study. Section 3 explains the materials, and Section 4 elaborates the proposed methods. Section 5 demonstrates the experimental results using quantitative comparisons, and discusses the observations. Section 6 concludes our studies and suggests a few recommendations for future research.

2. Background

Deep learning is a subset of a larger group of machine learning methods, which comprises of algorithms with hierarchical processing layers performing non-linear transformations to represent and learn data characteristics effectively [8]. Deep learning methods are currently being explored in various fields such as computer vision, audio and speech processing, natural language processing, information retrieval and bioinformatics [17]. Several deep learning architectures, for e.g., convolutional neural networks, deep belief networks and recurrent neural networks have been introduced [18], and have reported to achieve state-of-the-art results in a number of tasks. The goodness of data representation notably affects the performance of machine learning algorithms. Specifically, the existence of large-scale data has been recognized as a prerequisite for the success of many deep learning applications, leading to a convergence between the fields of deep learning and big data analytics [19].

One of the most frequently used methods of deep learning for two- dimensional data is the deep *convolutional neural networks*. These networks consist of interconnections emulating the arrangements in visual cortex of animals, where individual neurons are organized in a manner to respond to the overlapped tessellations comprising the visual field [20]. The principle transformations in deep CNNs consist mainly of combinations of convolu-

tional, pooling and fully connected layers, and their parameters are trained using backpropagation through these layers [21]. Deep learning using convolutional neural networks does not necessarily require prior computation of handcrafted features, and directly processes input images to compute self-derived and learned features during the training process. In recent years, CNNs have achieved breakthrough performance due to availability of large-scale training data and huge parallelization with graphics processing units (GPU) to speed up the application *i.e.* training and deployment process [22]. This has led to a development of number of time-efficient GPU-based frameworks [23] and rise of various deep learning algorithms for image analysis.

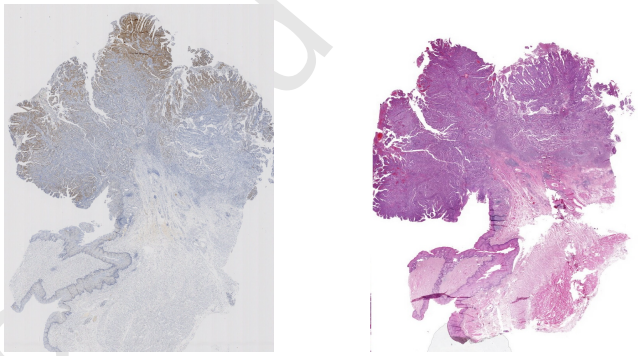
Random forest [24] is a popular ensemble learning method which constructs many decision trees and the final class is majority voted by the individual trees. Unlike deep CNN, training a random forest requires prior computation of suitable handcrafted features representing the image characteristics. In digital histopathology, the most commonly used handcrafted features [25], [26] are pixel-based (including texture, color and intensity), object-based (including structure and morphology) and architectural (including graph-based representations). Handcrafted features can be applied with random forests in a supervised learning approach for various classification and detection tasks such as [27] and [28]. Random forest has been selected for comparison with deep learning methods, because of high prediction accuracy among traditional machine learning algorithms and ability to efficiently handle large databases [24].

3. Materials

The gastric cancer dataset consists of whole slide images of surgical sections, each acquired from proximal or distal parts of stomach of a distinct patient, selected from an earlier study of 454 cases of gastric adenocarcinoma [29]. The slides were prepared using HER2 immunohistochemical staining and sections from the same tissue block were stained with haematoxylin and eosin (H&E) stain. The essential acquisition details about the studied samples are summarized in Table 1. The resulting H&E WSI have high contrast and visual quality, and are heterogeneous with variations in stain intensities, malignancy levels and inter-patient biological characteristics. Example of a WSI pair in HER2 and H&E stains is demonstrated in Figure 1.

Table 1: Acquisition Details

Acquisition Attributes	**Details**
Average thickness of sections	Approximately 4 μm
Staining methods	HER2 immunohistochemical stain, Haematoxylin and Eosin stain
WSI scanners	Leica SCN400 (HER2) and 3DHistech Panoramic-250 with extended depth of field [30] (H&E)
Sensor resolution	0.22 μm per pixel (at 40× objective magnification) with quadratic pixels
Number of annotated WSI	11 for cancer classification, 4 for necrosis detection
Underlying cancer grades (based on HER2 IHC)	0, 1+, 2+, 3+
Average WSI size (in pixels)	13.65 Gigapixels
Average number of pixels per annotation	HER2+ tumor: 231.7 $\times 10^6$ HER2- tumor: 1279.0$\times 10^6$ Necrosis: 48.01$\times 10^6$



(a) (b)

Figure 1: Examples of corresponding sections in (a) HER2 and (b) H&E stains.

All the available 11 WSI were utilized for our experiments in cancer classification, and four WSI were used in necrosis detection depending on the availability of labeled data. Ten expert pathologists previously marked polygon annotations in the HER2 WSI according to a 10% cutoff rule [29], depicting malignancy levels as HER2 negative areas (consisting of grades 0 and 1+) and HER2 positive areas (consisting of grades 2+ and 3+). Hence, to create the ground truth for cancer classification experiments, a semi-automatic

registration and annotation transformation procedure [5] was first applied to transform these polygon annotations from each HER2 WSI to the corresponding H&E WSI. To create the ground truth for necrosis detection experiments, annotations were marked by one expert pathologist directly in the H&E WSI indicating necrotic areas [7]. Data augmentation procedures are applied to generate a large number of image tessellations from the labeled WSI data, as described in detail in Section 4.2.

4. Methods

4.1. Schematic Overview

A schematic overview of the method is shown in Figure 2. It is further explained in the following sections.

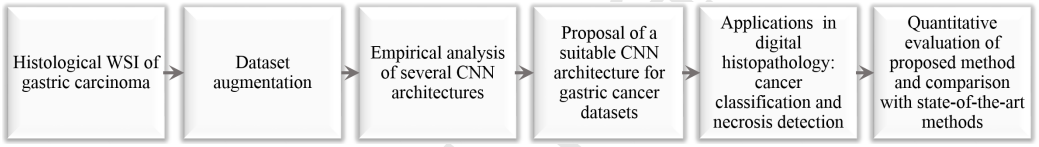


Figure 2: Schematic overview

4.2. Dataset Augmentation

In our studies, dataset augmentation was recognized as a prior requirement to expand a comparatively smaller amount of labeled WSI data, owing to many of the recent successes of deep learning algorithms with massive datasets [17]. However, the total number of instances in the large-scaled datasets is an experimental and task-specific question. In our study, standard data augmentation methods are used to generate large number of image tiles, each of size 512×512 pixels at highest magnification, from the WSI regions annotated by expert pathologists. These include overlapping by a factor of 0.3, and affine transformations such as rotations with 10 degree intervals, reflection, rotation after reflection and shear by a factor of 0.1. Brightness, contrast and intensity adjustments, and significant geometric deformations are not applied in order to preserve the salient texture, color and morphological properties of the original H&E stained tissue images. For cancer classification experiments, a total of 21,000 images from each slide are generated, thus, a large dataset containing 231,000 images is obtained. For necrosis detection, a smaller dataset is produced with a total of 47,130 images. The size of datasets depends on the availability of corresponding ground

truth marked by medical experts, especially in necrosis detection, maximum possible labeled data from the four slides is considered for experiments.

To speed up the reading process during training of convolutional neural networks, image datasets are first converted into lightning memory-mapped databases (LMDB) [31] storing images and corresponding labels. For our experiments, data augmentation is performed offline to organize the databases among cross validation rounds and to use them with traditional methods for comparative analysis. However, it is possible to perform this step online for speedup, such as in [9]. The process of dataset creation in our experiments is demonstrated in Figure 3, where an H&E WSI (same as Figure 1(b)) with five annotations of HER2 positive tumor marked by expert pathologists (one color per pathologist) in the corresponding HER2 WSI, is shown in Figure 3(a); a magnified region (5×) containing areas of agreement of most pathologists for the given class in Figure 3(b); and a few images after data augmentation in Figure 3(c).

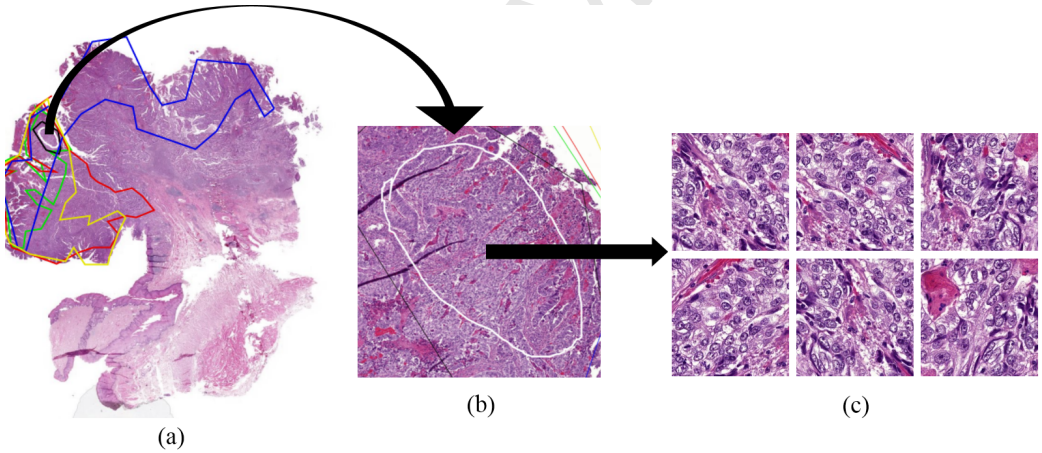


Figure 3: Example of H&E WSI with (a) a few annotations of HER2 positive tumor marked by expert pathologists in the corresponding HER2 WSI (b) a magnified (5×) region of agreement of most pathologists (c) example images after data augmentation at highest magnification (40×).

4.3. Empirical Analysis of CNN Architectures

Different CNN architectures are empirically studied to observe the behavior of variation in model characteristics (for e.g. network depth, layer properties, training parameters etc.) by training them from scratch on a

representative subset of the entire whole slide image data for cancer classification. For this purpose, a smaller dataset from five WSI is first created where image tiles are generated from three WSI for training the empirical CNN configurations, one WSI for validation phase (during training) and one WSI for test phase (during deployment). This empirical analysis acts as a *pilot study*, as it allows to establish the feasibility of deep learning methods for the described image analysis problems, considering a high visual complexity of gastric cancer histopathological images. It further facilitates basic understanding of the behavior of deep convolutional neural network architectures and economical utilization of time and space by first analyzing only a part of the available WSI image data.

Details of the most successful empirically evaluated CNN architectures are summarized in Table 2. The last row is AlexNet CNN [9] with the number of outputs in the last layer changed according to the cancer classification problem. There exist possibilities to extend the number of CNN architectures by using more combinations of the connected variables, however, performing an exhaustive evaluation is highly time consuming and restricted by the available hardware, so a few elementary CNN architectures are considered at this stage. Variants of the reported architectures are also tested without further improvements. On observing the classification performance using training curves, overall validation and test accuracies, the highlighted CNN architecture with nine layers (three convolutional layers, three pooling layers and three fully connected layers) achieves favorable results for all the three malignancy levels on the small representative dataset, that will tentatively improve by using more abundant examples. Based on empirical analysis and insight to accurately model the characteristics of histological images, this self-designed CNN architecture is selected for further evaluation with entire available WSI data. Later, the proposed CNN is also applied for necrosis detection.

We find that for our gastric cancer images in general, convolution layers with larger kernels followed by smaller ones lead to a better training procedure with a lower training loss and higher accuracies. Similar relative convolutional kernel sizes are adapted by AlexNet and other popular CNN architectures. Additionally, an architecture with smaller convolutional kernels followed by larger ones was also tested, but it produced reasonable accuracies for only two classes and completely failed to classify the HER2 negative tumor class, indicating the importance of larger neighborhoods in histopathological images. Poor performance was also observed in another model with hinge

Table 2: Details of most successful empirically evaluated CNN architectures for cancer classification on gastric cancer representative datasets

Model number	**Total number of layers**	**Number of convolution layers**	**Kernel sizes: convolution layers**	**Number of feature maps**	**Kernel sizes: pooling layers**	**Number of fully connected layer outputs**	**Average multi-class accuracy (cancer classification)**	
							validation WSI	**test WSI**
1	6	2	3, 3	16, 16	3, 3	128, 3	0.3235	0.3333
2	6	2	7, 3	16, 16	2, 2	256, 4	0.5208	0.5140
3	7	2	9, 9	16, 16	3, 3	128, 128, 3	0.6025	0.5417
4	7	2	7, 5	16, 16	2, 2	256, 128, 3	0.6000	0.5432
5	9	3	7, 5, 3	24, 16, 16	2, 2, 2	256, 128, 3	0.6056	0.5571
6	10	4	9, 7, 5, 3	32, 128, 128, 128	3, 3, 3	2048, 2048, 3	0.5951	0.4376
7	11	5	11, 5, 3, 3, 3	96, 256, 384, 384, 256	3, 3, 3	4096, 4096, 3	0.3948	0.3948

loss function, whereas softmax loss is used in the architectures mentioned in Table 2, thus, found more suitable for learning. Reasons for testing moderately deep networks in our experiments are the limited availability of training data (especially for necrosis detection) and system memory. Higher network depths would cause a rise in the number of network parameters leading to non-optimal learning with smaller datasets (overfitting), and also result in higher space and time complexity. However, a more detailed study of the CNN architectural design to determine the relationship between design parameters is a future direction of our research, and can plausibly lead to an improvement of results obtained so far in our experiments.

Another option was to fine-tune a pre-trained CNN, but was intuitively skipped due to the following reasons. As the name suggests, fine-tuning of a pre-trained network is generally performed when the network trained successfully with visually very similar images. For example, AlexNet [9] was

originally tested on general object categorization with ILSVRC dataset [32] and was later fine-tuned for Flickr style images [33]. Such fine-tuning doesn't appear useful for our complex histological WSI data. Moreover, classification results after training AlexNet from scratch during empirical analysis on a smaller representative dataset are unexceptional. Therefore, for the purpose of completeness, the performance of AlexNet and proposed CNN architecture are comparatively evaluated using the same cross validation strategies for both approaches and training from scratch on the entire available WSI data.

4.4. Proposed Convolutional Neural Network Architecture

The selected self-designed CNN architecture is a purely supervised feed forward network shown in Figure 4. It consists of nine stages, where the first six stages comprise of convolution and pooling operations, and the last three stages are fully connected layers. For the input images, a batch size of 8 is used during training phase and 10 during deployment phase to improve efficiency using batch processing. Image mean values are calculated and subtracted from the images during both phases, such as in the AlexNet framework [9]. The detailed descriptions of the various layers are as follows.

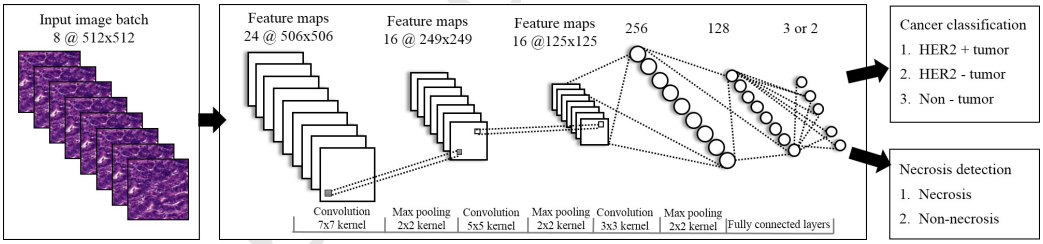


Figure 4: Proposed CNN architecture

4.4.1. Convolutional Layer

In each convolutional layer, first the output of previous layers is convolved with multiple learned weight matrices called filter masks or learned kernels. Then the result is processed by a non-linear operation to generate the layer output. The linear operation in the k^{th} convolution layer, whose input is denoted by x^{k-1} (output of the $[k-1]^{th}$ layer) comprises of a two-dimensional

convolution [34] as shown in Equation 1.

$$o^k[m,n] = x^{k-1}[m,n] * W^k[m,n] + b_k = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} x^{k-1}[u,v] W^k[m-u, n-v] + b_k \tag{1}$$

where, W^k denotes the learned weight matrix and b_k denotes bias or offset of the k^{th} convolution layer. For our experiments, three convolutional layers are used with filter sizes 7×7, 5×5 and 3×3 and 24, 16 and 16 feature maps respectively. The size of learned filters should be similar to the size of patterns to be detected. This decision is task dependent and in our classification problems, subtler details in tissue regions need to be resolved, hence, smaller filter sizes are selected. Further, the weight initialization is performed using Gaussian functions with standard deviation 0.01, and bias is kept constant at 0.1.

The outputs o^k are applied to an operation which improves the learning process by increasing non-linear properties of the network, and rectified linear units (ReLU) are selected due to benefits explained in [9]. These are non-saturating activation functions where piece-wise linear tiling can be achieved. The basic ReLU operation is given by Equation 2. The cascade of operations in the convolutional layer are depicted in Figure 5.

$$x^k = max(0, o^k) \tag{2}$$

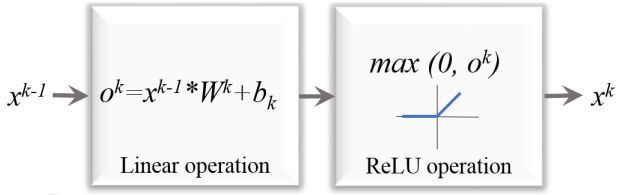


Figure 5: Cascade operations in the k^{th}convolutional layer

4.4.2. Pooling Layer

A pooling layer performing *max-pooling* [9] is applied after each convolutional layer. Max-pooling involves splitting the filter output matrix into non-overlapping grids and taking the maximum value in each grid as the value in the reduced matrix. Hence, it combines responses at different locations and adds robustness to small spatial variations, thereby increasing translational invariance, along with reducing spatial resolution. We use a kernel size 2×2 and stride two for maxpooling operation after each convolutional layer.

4.4.3. Fully-connected Layer

In the fully-connected layer, neurons are connected to the neurons in previous and next layers, such as in conventional neural networks [35]. In the proposed CNN, three fully connected layers are used with 256, 128 and three (or two) outputs respectively, as the number of outputs in last layer depends on the classification problem at hand. Bias is set to 0.01.

Dropout method [36] assists in reducing overfitting, especially when the available training data is limited such as the WSI data. During each iteration, individual nodes along with incoming and outgoing edges are removed from the network, and are later returned along with their initial weights. In our approach, after each of the first two fully-connected layers, the dropout ratio *i.e.* the probability of dropping any input for both stages is set to 0.25.

4.4.4. Learning Properties

Learning of the convolutional neural network is based on measuring a loss function (also called objective function, error function, cost function) that indicates the error of learned network parameters. The learning objective is to compute the parameters to minimize the loss function. Softmax function [37] (Equation 3) is the probability of class c_i given input X, where z_i represents the score for i^{th} class among total C classes. The softmax loss E is calculated as negative log likelihood of the softmax function (Equation 4), where N denotes the length of the class vector.

$$P(c_i|X) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{3}$$

$$E = -\frac{1}{N}\sum_{n=1}^{N} log(P(c_n|X)) \tag{4}$$

The method used for optimizing the loss minimization is called *Stochastic Gradient Descent with Momentum* [38], [39]. For the t^{th} iteration, the update process is denoted by Equation 5 [23], [39], where θ_t denotes the current weight update, θ_{t-1} is the previous weight update, w represents the weights, E is the average loss over the dataset, $\nabla E(w)$ is negative gradient, η is the learning rate and $\mu \in [0, 1]$ represents momentum used for speeding up the convergence of gradient and preventing oscillations.

$$\begin{aligned} \theta_t &= \mu\theta_{t-1} - \eta\nabla E(w_{t-1}) \\ w_t &= w_{t-1} + \theta_t \end{aligned} \tag{5}$$

In the learning phase, the two hyperparameters η and μ are required to be experimentally determined. After the empirical analysis stage, η has been set to 0.001 initially, and is decreased by a factor of 0.1 after every 20,000 iterations, and μ is fixed at 0.9. Around eight training epochs are performed during each training phase after initial experimental analysis of training curves on our datasets.

4.5. Implementation Details

Training and testing of the convolutional neural networks is performed using GPU programming to achieve higher time efficiency over CPU [22]. The available hardware is NVIDIA GeForce GTX 660 with 2GB memory. Caffe [23] is an open-source C++ based library with Python and Matlab bindings for training and deploying CNNs, and also contains reference models such as AlexNet [9] for experimentation. Conversion of image datasets into LMDB, computations of image means and implementation of the proposed CNN architecture are performed using the modules in Caffe framework. The reference implementation of AlexNet is directly utilized from Caffe. The operating system is Ubuntu 14.04 and system specifications are Intel Core i7-4790 processor at 3.60 GHz with 15 GB RAM. On implementing the proposed CNN architecture on the GPU, each training iteration takes *ca.* 0.7 seconds, hence, a training phase requires *ca.* 40.8 hours for cancer classification and *ca.* 11.7 hours for necrosis detection. Deployment of the CNN requires around 0.25 seconds per image for the described experimental setup.

Dataset creation, preprocessing steps, comparative evaluation using handcrafted features and random forest classification are performed using another computer with Windows® operating system, Intel Core i7-3700 processor at 3.40 GHz with 16 GB RAM. Whole slide image data is accessed in C# using the VMscope software support [40] to generate the image dataset for our experiments. For the data augmentation part, .NET platform is used with C# implemented modules. Accord [41] libraries are employed for feature extraction of GLCM measures and LBP histograms. AForge framework utilities [42] are used to generate RGB and gray histograms in timely manner. HSV histograms are computed using the EmguCV wrapper [43] of the OpenCV library in C#. Python modules are applied from scikit-image [44] for computing Gabor filter-bank features, and from scikit-learn [45] for random forest machine learning and classification.

5. Experimental Results

5.1. Applications

As stated earlier, deep learning using convolutional neural networks is explored for two computer-based classification applications, namely, cancer classification based on IHC and necrosis detection.

5.1.1. Cancer Classification based on IHC

H&E stain is routinely used for primary diagnosis in histopathology, and immunohistochemical staining is subsequently recommended to reveal details of malignancy. Immunohistochemical staining generally involves costlier preparation, especially the HER2 IHC stain was recently introduced as a biomarker for gastric cancer [29]. Therefore, H&E stained WSI are analyzed in our study due to lower cost and wider usage. Pathologists can visually recognize the different malignancy levels by observing HER2 stained slides using optical or virtual microscopy methods but require a greater time and effort to identify the corresponding tumor areas in H&E stain. This work explores the possibility of presence of subtle differences in tissue properties between malignancy levels in the H&E stained tissue regions, which are difficult to visually identify but can be detected by computer-based image analysis methods.

5.1.2. Necrosis Detection

This application attempts to differentiate between image regions in order to find the existence of tumor necrosis in the H&E stained WSI. Due to a distinct appearance of tumor necrosis compared to the living tissue, automatic pattern recognition methods can be suitably applied for detecting necrotic areas in histopathological images. It may be harder to reach diagnostic conclusions by observing a WSI containing necrotic regions, and manual identification of necrosis using visual inspection can be a time-consuming task. Automatic necrosis detection can provide useful information about the the type and extent of malignancy and help in formation of prognosis, as higher necrosis may promote tumor growth and consequently lead to a lower possibility of survival [46]. In some cancer patients, (neoadjuvant) chemotherapy is followed by surgery and histological diagnosis, where determining the extent of necrosis can prove useful. Moreover, the detected necrotic areas can be excluded in order to carefully analyze the remaining living tissue. Therefore, necrosis detection can constitute a preparatory stage that is helpful to

pathologists and subsequent analysis tools for more precise disease observation and characterization.

5.2. Comparison with State-of-the-Art Methods

Performance of the selected CNN architecture is compared with state-of-the-art approaches for image analysis in digital histopathology. These include handcrafted texture and color descriptors such as the GLCM features, Gabor filter-bank features, LBP histograms, gray histograms, HSV histograms and RGB histograms followed by random forest machine learning. It is also compared with the well-established AlexNet CNN framework.

In literature, the well-known handcrafted features include texture, color and intensity, morphological and architectural measurements and have been extremely successful for image analysis applications in digital histopathology, for e.g. [47], [48], [49], [50], [51], [52], [6], [53]. Some of these have been evaluated in our experiments to measure comparison with deep learning methods. Statistical descriptors derived from the gray level co-occurrence matrix (GLCM) include the 14 texture features described in [54]. Gabor filter-bank functions are similar to two dimensional receptive fields of human visual system [55], hence, a set of 32 Gabor filter-bank features is computed to describe texture. Local binary patterns (LBP) operator thresholds the neighborhood of each pixel to generate a binary code [56] and the corresponding labels are used to obtain LBP histograms of 256 features. Additionally, gray histograms, HSV histograms and RGB histograms are calculated for each image to yield 256, 692 and 768 length feature vectors respectively. These histograms can accurately depict the intensity or color distributions in H&E stained histological images using gray levels, HSV and RGB color spaces [57]. The configuration of random forest classifiers consists of 1000 trees for each classification and square root of total number of features for best split, initially estimated from a predefined range by out-of-bag error stabilization [45].

The AlexNet framework is one of the most widely accepted deep CNN algorithms due to its tremendous accomplishment in general image categorization [9], thus, nowadays considered as a baseline for comparison of deep learning algorithms. Gastric cancer histological datasets are also tested using AlexNet for both applications. Input images used for training and classification are cropped to size 227×227 from the center, similar to the images in AlexNet framework. Number of training rounds and batch sizes are kept constant for the two types of deep learning approaches.

5.3. Quantitative Evaluation

For quantitative evaluation, two cross validation strategies, namely, *k-fold stratified shuffle split* and *leave-a-patient-out* are used to compare performance of different approaches and establish our findings. In the *k-fold stratified shuffle split* evaluation method where k=3, two-third of the whole dataset is used for training and remaining one-third for classification in each round, where the training and classification datasets are randomly selected without any overlap with equal number of samples from each class. In the *leave-a-patient-out* strategy, in each round, the samples belonging to one patient are excluded from the training dataset and used for testing the classification method. This cross validation strategy helps to study the practical usability of the described methods, because of variations between patients owing to a heterogeneous nature of our dataset.

The average multi-class classification accuracy using the proposed CNN architecture and state-of-the-art methods is shown in Table 3 for cancer classification and Table 4 for necrosis detection, depicting the overall performance of handcrafted features with traditional machine learning versus deep learning algorithms. Corresponding plots for individual classes are given in Figure 6 and Figure 7, showing the distributions of results among cross validations for cancer classification and necrosis detection respectively. Examples of learning curves of the proposed CNN architecture are demonstrated in Figure 8 for both applications and randomly selected training round of each of the cross validation strategies.

Table 3: Average classification accuracy of applied methods for cancer classification

Methods	**HER2+ tumor**	**HER2- tumor**	**Non-tumor**	**Overall**
Handcrafted Features and Random Forests				
GLCM features	0.6893	0.5945	0.7989	0.6942
Gabor filter-bank features	0.6527	0.5548	0.8193	0.6756
LBP histograms	0.6479	0.5477	0.6789	0.6248
Gray histograms	0.7055	0.6382	0.8775	0.7404
HSV histograms	0.7062	0.6604	0.8672	0.7446
RGB histograms	0.7467	0.6670	0.8785	0.7641
Convolutional Neural Networks				
AlexNet CNN	0.7533	0.6613	0.7837	0.7328
Proposed CNN architecture	**0.6959**	**0.5809**	**0.8203**	**0.6990**

Table 4: Average classification accuracy of applied methods for necrosis detection

Methods	Necrotic	Non-necrotic	Overall
Handcrafted Features and Random Forests			
GLCM features	0.5458	0.7486	0.6472
Gabor filter-bank features	0.7427	0.7134	0.7280
LBP histograms	0.7123	0.6421	0.6772
Gray histograms	0.7290	0.6392	0.6841
HSV histograms	0.7262	0.6929	0.7096
RGB histograms	0.6678	0.6769	0.6723
Convolutional Neural Networks			
AlexNet CNN	0.5847	0.7444	0.6646
Proposed CNN architecture	**0.7718**	**0.8570**	**0.8144**

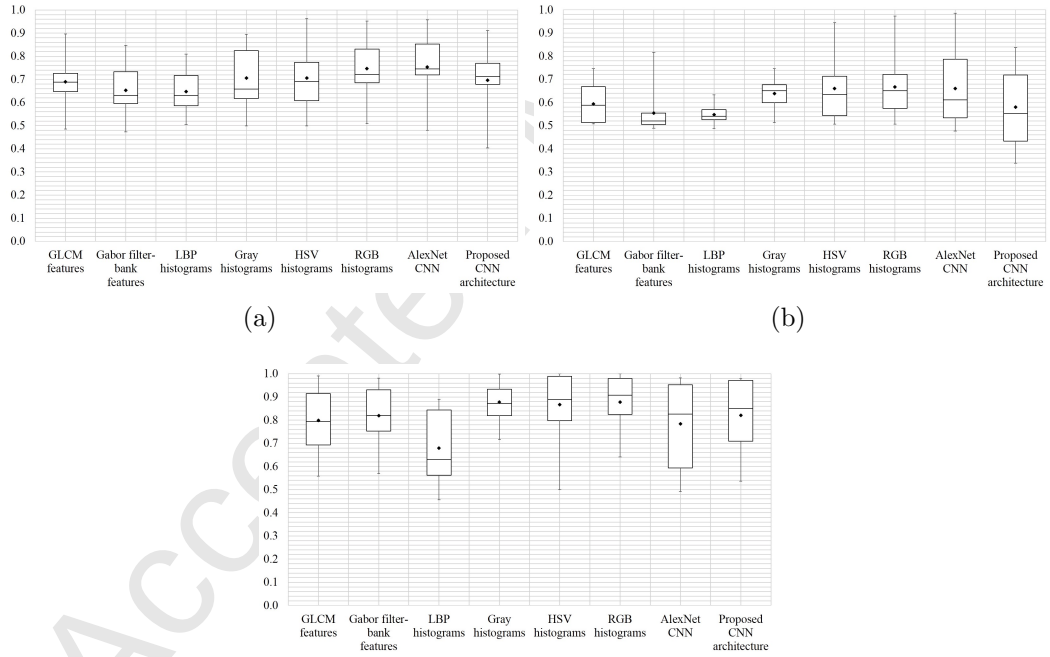


Figure 6: Results of cross validation rounds for cancer classification experiments for classes (a) HER2+ tumor (b) HER2- tumor (c) Non-tumor.

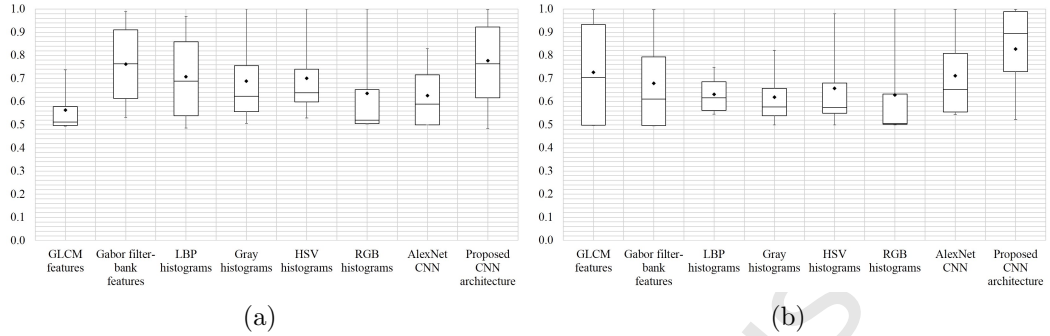


Figure 7: Results of cross validation rounds for necrosis detection experiments for classes (a) Necrotic (b) Non-necrotic.

5.4. Observation and Discussion

From the experimental results, we observe the following. For cancer classification, using all the described methods, non-tumor class is classified with best accuracy, followed by HER2 positive tumor, and then HER2 negative tumor. The results indicate that non-tumor areas are visually distinguishable, hence, comparatively simpler to discriminate by the classifiers from the cancer-affected tissue. HER2 negative tumor has the least accurate classification, because it can be described as the beginning of tumor proliferation and as an intermediate stage between non-tumor and higher malignancy (HER2 positive tumor). It consists of the tissue regions where tumor is partially developed, so there is less or no immunohistochemical response from cells lying in these regions in the HER2 stained tissue. Resolving this confusion proves as the main challenge to obtain good identification rates for both malignancy levels in H&E stained tissue images. This effect was also observed using handcrafted graph-based features and other classification methods in [6] with the help of confusion matrices for the two types of cross validations, and similar confusion matrices have also been generated for this work showing comparable trends for the three classes. Furthermore, it can be observed from the results that among feature extraction methods, RGB histograms with random forests show the most satisfactory performance followed by HSV and gray histograms. Both the CNN architectures compare favorably to most handcrafted features and traditional machine learning, however, their performance is slightly lower than three types of histograms. It may lead to the

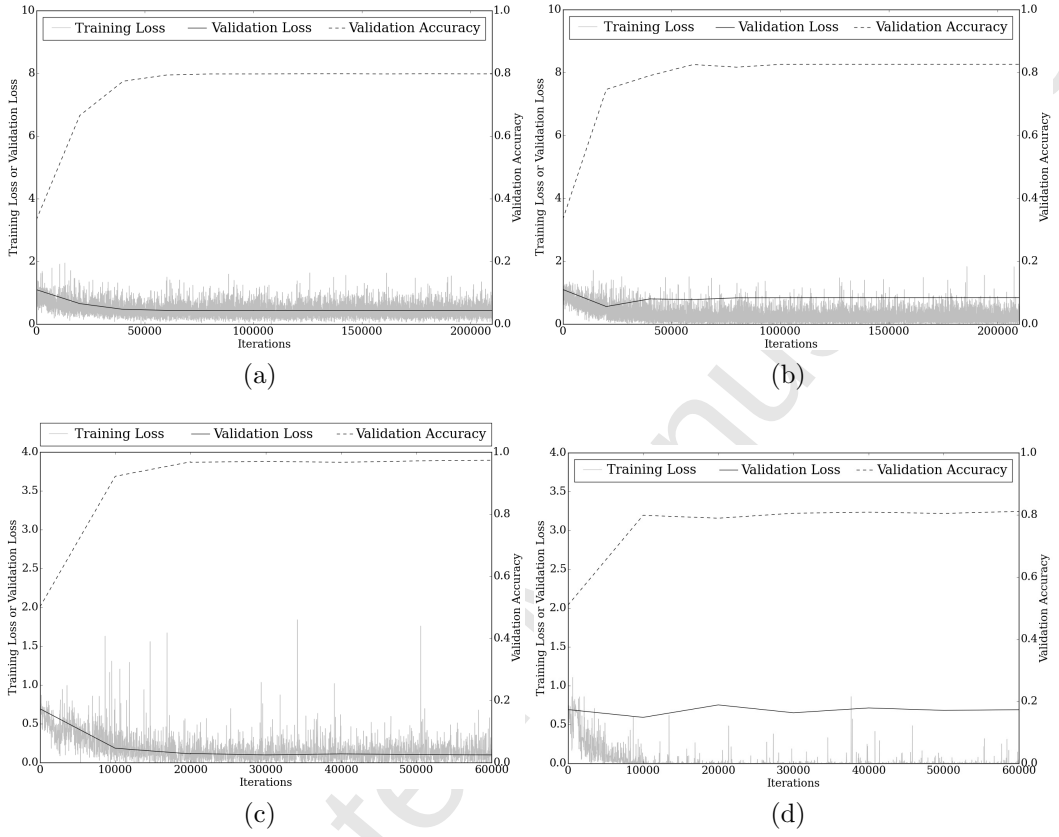


Figure 8: Examples of learning curves of random training rounds using the proposed CNN architecture for (a) cancer classification in *k-fold stratified shuffle split* cross validation (b) cancer classification in *leave-a-patient-out* cross validation (c) necrosis detection in *k-fold stratified shuffle split* cross validation (d) necrosis detection in *leave-a-patient-out* cross validation.

conclusion that the H&E stain distribution is most beneficial for representing discriminating information to classify gastric cancer tissue images into one of the three malignancy levels using random forest classifiers. Moreover, we note that using the selected CNN architecture, accuracies are higher when trained with the entire available WSI dataset compared to the initial empirical experiments, and similar phenomenon is observed for AlexNet framework. This behavior was expected because of subsequent use of expanded datasets built from the entire available WSI data, hence, the networks are trained with

large-scale visual information with diverse examples to classify the unknown images more effectively. On comparing the performance of the two CNN architectures, it is found that AlexNet has marginally superior detection rates. However, it should be emphasized that, during the *k-fold stratified shuffled split* cross validation, AlexNet is more accurate (0.9682) than the proposed CNN (0.8328) whereas *leave-a-patient-out* cross validation leads to an opposite observation (0.4973) for AlexNet versus the proposed CNN (0.5653), yielding a slightly higher overall accuracy. A better *leave-a-patient-out* performance is a desirable characteristic reflecting the robustness of the method for practical application in a clinical setting where unknown samples can be automatically classified with trained networks of previous cases, and this requirement is better achieved using the proposed CNN.

On observing the results for necrosis detection, we find that the proposed CNN architecture has the best overall rates (0.8144) and outperforms all the hand-engineered traditional methods and also the AlexNet deep CNN for both classes, hence, the most desirable performance is achieved using the proposed CNN architecture. It is also interesting as the number of images required for training the deep convolutional models was lower compared to cancer classification problem, which indicates necrosis to be visually simpler to detect due to its higher distinguishability. However, sometimes more variation is observed between the two types of cross validations, as compared to cancer classification, indicating the requirement of more diverse and comprehensive databases during the training phase.

In general, for both cancer classification and necrosis detection, more favorable performance is achieved using *k-fold stratified shuffled split* compared to *leave-a-patient-out* cross validation. The observed results using the former technique ensure highly desirable outputs in scenarios where partially annotated data in single or multiple large-sized WSI of the same patient is available. In contrast, leave-a-patient-out performance indicates the biological variability between limited number of available cases, as a few difficult ones are observed to show poor recognition rates and affect the overall classification measurements of the proposed system. It can be emphasized that the extension of current ground truth data with more patients and lower number of images per patient for training, can provide a more robust classification algorithm. In this way, it may be more feasible to evaluate the approach purely on the basis of the second type of cross validation that will be explored in our future studies.

The improvement of performance of the proposed CNN architecture over

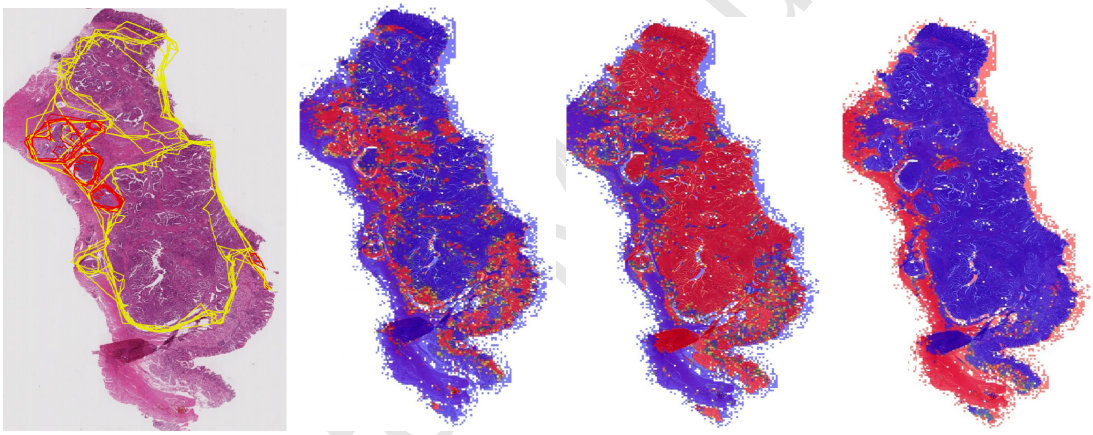
AlexNet during *leave-a-patient-out* cross validation in cancer classification, and during both cross validation strategies in necrosis detection can be mostly attributed to its properties, leading to a more effective representation of subtler details with higher robustness to inter-patient variations, which may not be in the case for AlexNet. Additionally, the proposed CNN architecture uses larger regions compared to AlexNet, which is considered as an advantage because context and neighborhood are important characteristics in histological images and captured more efficiently by considering higher resolutions. Training of AlexNet was also attempted using the same configuration and image sizes, however, the specified GPU was unable due to higher memory requirements.

The plots in Figure 6 and Figure 7 are interesting in order to investigate the behavior of classification methods over multiple cross validation rounds. These plots reinforce that the proposed CNN architecture performs reasonably well compared to other tested methods in cancer classification, and outperforms them in necrosis detection. An encouraging observation is that the CNN has comparatively stable performance with lower inter-round variations, except for HER2 negative tumor class, which is not surprising due to its high visual complexity. Certain handcrafted features like Gabor filter-bank responses and LBP histograms show more unpredictable behavior. AlexNet shows performance variability for classes such as non-tumor, which is easily distinguishable by other methods.

Lastly, we observe the learning curves in Figure 8 generated by few rounds for both the problems, for the given number of epochs. It can be seen that that they are generally smooth with a declining training error and increasing validation accuracy with iterations, that become constant after several iterations. In Figure 8(d), the validation loss is nearly constant due to characteristics of validation data, but training loss decreases and validation accuracy increases to become constant, as desired for successful training.

A probability map for each category in the two classification problems can be generated for a WSI for visualization purposes. During classification of an image tile, the prediction probability estimate of each class can be expressed using a color scale from blue to red representing values in range [0,1]. The probability representations of constituent non-overlapping image tiles overlay the original H&E WSI to generate a probability map for each class. An example for cancer classification of one of the H&E WSI is shown in Figure 9. Figure 9(a) depicts the original WSI with pathologists' annotations showing HER2 positive areas with red and HER2 negative areas with yellow

polygons. Figures 9(b), 9(c) an 9(d) are the probability maps for HER2 positive, HER2 negative and non-tumor. Because of very high resolution of the WSI, the constituent image tiles appear as small colored pixels at a lower magnification (0.3×). The results obtained are mostly consistent with the expert annotations made in the corresponding HER2 WSI, indicating desirable behavior of the proposed CNN method for analysis of H&E stained gastric cancer WSI based on their HER2 immunohistochemistry. Higher confusion and prediction error is observed mainly in the lower right part of the WSI, especially in the tumor classes.



(a) (b) (c) (d)

Figure 9: Illustrative examples of classification result (a) Original H&E WSI with pathologists' annotations for cancer classification based on IHC, and corresponding probability maps using proposed CNN architecture for (a) HER2 positive tumor (b) HER2 negative tumor (c) non-tumor at a low magnification (0.3×).

6. Conclusions and Recommendations

In this paper, whole slide images of gastric carcinoma are analyzed with the help of deep learning methods in digital histopathology. Our previous studies on H&E stained gastric carcinoma images in [6] and [7] involve graph-based analysis and AdaBoost classification based on their HER2 immunohistochemistry, and necrosis detection using texture-based approaches and SVM classification respectively. This study addresses these histological image analysis tasks using deep convolutional neural networks. Firstly, several

CNN architectures are empirically investigated to determine a suitable design for learning significant characteristics of histological images. A preliminary CNN architecture is proposed and applied for two classification problems, namely, cancer classification based on immunohistochemical response and necrosis detection. The performance of the designed CNN is quantitatively evaluated and compared with state-of-the-art handcrafted features with random forest classification. It is also compared to the AlexNet CNN framework which is popular in deep learning for general object classification.

In general, deep learning methods compare favorably to traditional methods using hand-engineered features and random forest machine learning. Furthermore, the self-designed CNN architecture shows promising results compared to AlexNet CNN framework. For cancer classification based on IHC, non-tumor can be easily distinguished from tumors, but it is more difficult to distinguish the two types of tumor classes. RGB histograms show highly desirable performance, however, both the studied convolutional neural networks also perform reasonably well. For necrosis detection, the proposed CNN is trained using a smaller image dataset and has the best performance over all the traditional methods and the AlexNet framework. Hence, it can be concluded that, for both classification applications, the proposed CNN architecture adequately represents H&E stained histopathological images of gastric cancer at high magnification with varying stain intensities and malignancy levels.

A few limitations of our method are highlighted as follows. Firstly, like all other deep learning applications so far, our method also requires training with large-scale datasets containing thousands of images. This problem is currently addressed using data augmentation strategies on the available datasets, but it will be a more appropriate direction to expand the available ground truth by increasing the number of cases. For our experiments, the ground truth data was generated by expert pathologists by marking manual annotations in WSI which was a laborious task demanding considerable time, hence, our studies are currently limited to smaller datasets. Also, training the proposed CNN from scratch, even with a fast (although not optimized) implementation on GPU, requires around two days for the current configuration. On the other hand, an optimized implementation of feature extraction and random forest machine learning requires relatively smaller duration to complete training for same sized data. This trade-off between accuracy versus time efficiency requires further experimental evidence to ensure an upper hand of deep learning algorithms in digital histopathology. Another chal-

lenge was the empirical analysis of CNN architectures to determine an optimal framework for histological image analysis of gastric cancer and currently this stage is based on cancer classification. The selected CNN architecture is applied for detecting necrosis to explore the generalizability of CNNs versus handcrafting features for individual problems, and our approach shows a reasonable performance in both applications. Nevertheless, other possibilities can be studied in future which were limited now due to time and hardware requirements.

The next steps of this study will be to enhance and refine the designed CNN architecture in order to reduce confusion by achieving higher classification accuracy for the mentioned classification problems, and to optimize program execution. Specifically, cancer classification based on IHC requires refinement in categories and classification strategy for improved performance, and hierarchical classification is being currently explored for the same. Moreover, as stated above, the methods have currently been tested on smaller number of WSI which are being expanded to more patients along with annotations. The proposed CNN architecture will also be applied on other diverse classification problems in histology, to further investigate its generalizability. Other interesting directions include comparative evaluation with other successful CNN architectures in digital histopathology, and ensemble learning [58], [59] to combine classifiers based on traditional machine learning and deep learning in order to harmonize their individual strengths. The comprehensive aim of our studies is to achieve effective computer-based histological image analysis using the routinely examined H&E stained WSI of gastric cancer.

Acknowledgements

This work is financially supported by the German Academic Exchange Service (DAAD). We are grateful to the Department of Pathology, Christian-Albrechts University, Kiel, Germany, especially Dr. Christine Böger and Mr. Hans-Michael Behrens for providing the gastric cancer WSI specimens with pathologists' annotations for cancer classification and important medical knowledge as required for this study. We express deep gratitude towards Dr. Olaf Ronneberger, Institut für Informatik, Albert-Ludwigs-Universität, Freiburg, Germany and Dr. Grégoire Montavon, Machine Learning Group, Technical University Berlin, Berlin, Germany for sharing their knowledge about the important concepts of deep learning through lectures and discus-

sion respectively, which laid a strong foundation for this work. We thank Mr. Sebastian Lohmann, Charité University Hospital, Berlin, Germany, for contributing in the Gabor filter-bank feature extraction and Mr. Björn Lindequist, Charité University Hospital, Berlin, Germany, for his contribution in the pre-processing stage. We sincerely thank VMscope GmbH for providing suitable software tools to access whole slide image data as required for our work. We are thankful to organizers of ECDP 2016 for making the audio web-cast of the corresponding presentation available at [60].

References

[1] Rugge M, Fassan M, Graham DY. Epidemiology of Gastric Cancer. In: Strong VE, editor. Gastric Cancer. Springer International Publishing. ISBN 978-3-319-15825-9; 2015, p. 23–34. URL: http://dx.doi.org/10.1007/978-3-319-15826-6_2. doi:10.1007/978-3-319-15826-6_2.

[2] Ficsor L, Varga V, Berczi L, Miheller P, Tagscherer A, Wu MLc, et al. Automated virtual microscopy of gastric biopsies. Cytometry Part B: Clinical Cytometry 2006;70(6):423–31.

[3] Zaitoun A, Al Mardini H, Record C. Quantitative assessment of gastric atrophy using the syntactic structure analysis. Journal of clinical pathology 1998;51(12):895–900.

[4] Cosatto E, Laquerre PF, Malon C, Graf HP, Saito A, Kiyuna T, et al. Automated gastric cancer diagnosis on H&E-stained sections; ltraining a classifier on a large scale with multiple instance machine learning. In: SPIE Medical Imaging. International Society for Optics and Photonics; 2013, p. 867605–.

[5] Sharma H, Zerbe N, Heim D, Wienert S, Behrens HM, Hellwich O, et al. A Multi-resolution Approach for Combining Visual Information using Nuclei Segmentation and Classification in Histopathological Images. In: Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISAPP 2015). Scitepress. 2015, p. 37–46.

[6] Sharma H, Zerbe N, Heim D, Wienert S, Lohmann S, Hellwich O, et al. Cell nuclei attributed relational graphs for efficient representation and

classification of gastric cancer in digital histopathology. In: SPIE Medical Imaging. International Society for Optics and Photonics; 2016, p. 97910X–.

[7] Sharma H, Zerbe N, Klempert I, Lohmann S, Lindequist B, Hellwich O, et al. Appearance-based Necrosis Detection Using Textural Features and SVM with Discriminative Thresholding in Histopathological Whole Slide Images. In: Bioinformatics and Bioengineering (BIBE), 2015 IEEE International Conference on. IEEE; 2015, p. 1–6.

[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012, p. 1097–105.

[10] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation 1989;1(4):541–51.

[11] Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: SPIE Medical Imaging. International Society for Optics and Photonics; 2014, p. 904103–.

[12] Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer; 2013, p. 411–8.

[13] Ronneberger O, P.Fischer , Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI); vol. 9351 of *LNCS*. Springer; 2015, p. 234–41. URL: `http://lmb.informatik.uni-freiburg.de//Publications/2015/RFB15a`; (available on arXiv:1505.04597 [cs.CV]).

[14] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial

and stromal regions in histopathological images. Neurocomputing 2016;191:214 –23. URL: http://www.sciencedirect.com/science/article/pii/S0925231216001004. doi:http://dx.doi.org/10.1016/j.neucom.2016.01.034.

[15] Hou L, Samaras D, Kurç TM, Gao Y, Davis JE, Saltz JH. Efficient Multiple Instance Convolutional Neural Networks for Gigapixel Resolution Image Classification. arXiv preprint arXiv:150407947 2015;URL: http://arxiv.org/abs/1504.07947.

[16] Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014, p. 806–13.

[17] Deng L, Yu D. Deep learning: methods and applications. Foundations and Trends in Signal Processing 2014;7(3–4):197–387.

[18] Deng L. Three classes of deep learning architectures and their applications: a tutorial survey. APSIPA transactions on signal and information processing 2012;.

[19] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of Big Data 2015;2(1):1–21.

[20] Hinton GE. What kind of graphical model is the brain? In: IJCAI; vol. 5. 2005, p. 1765–75.

[21] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. 2012, p. 3642–9. doi:10.1109/CVPR.2012.6248110.

[22] Strigl D, Kofler K, Podlipnig S. Performance and scalability of GPU-based convolutional neural networks. In: 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing. IEEE; 2010, p. 317–24.

[23] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia. ACM; 2014, p. 675–8.

[24] Breiman L. Random forests. Machine learning 2001;45(1):5–32.

[25] Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. Journal of the American Medical Informatics Association 2013;20(6):1099–108.

[26] Sharma H, Zerbe N, Lohmann S, Kayser K, Hellwich O, Hufnagl P. A review of graph-based methods for image analysis in digital histopathology. Diagnostic Pathology 2015;1(1).

[27] DiFranco MD, OHurley G, Kay EW, Watson RWG, Cunningham P. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. Computerized medical imaging and graphics 2011;35(7):629–45.

[28] Sommer C, Fiaschi L, Hamprecht FA, Gerlich DW. Learning-based mitotic cell detection in histopathological images. In: Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE; 2012, p. 2306–9.

[29] Behrens HM, Warneke VS, Böger C, Garbrecht N, Jüttner E, Klapper W, et al. Reproducibility of Her2/neu scoring in gastric cancer and assessment of the 10% cut-off rule. Cancer medicine 2015;4(2):235–44.

[30] Bradley AP, Wildermoth M, Mills P. Virtual microscopy with extended depth of field. In: Digital Image Computing: Techniques and Applications, 2005. DICTA'05. Proceedings 2005. IEEE; 2005, p. 35–.

[31] Chu H. Mdb: A memory-mapped database and backend for openldap. LDAP11 2011;.

[32] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 2015;115(3):211–52.

[33] Xia Y. Fine-tuning for Image Style Recognition. 2015.

[34] Podlozhnyuk V. Image convolution with CUDA. NVIDIA Corporation white paper, June 2007;2097(3).

[35] Kröse B, van der Smagt P. An introduction to neural networks. 1993.

[36] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 2014;15(1):1929–58.

[37] Bishop CM. Pattern Recognition and Machine Learning. Springer, New York; 2006.

[38] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th international conference on machine learning (ICML-13). 2013, p. 1139–47.

[39] Bottou L. Stochastic gradient descent tricks. In: Neural Networks: Tricks of the Trade. Springer; 2012, p. 421–36.

[40] VMscope GmbH . VMscope Products. `http://www.vmscope.com/produkte.html`; 2010.

[41] Souza CR. The Accord.NET Framework. `http://www.accord-framework.net`; 2014.

[42] Kirillov A. Aforge.net framework. `http://www.aforgenet.com`; 2013.

[43] Shi S. Emgu CV Essentials. Packt Publishing Ltd; 2013.

[44] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: Image processing in Python. PeerJ 2014;2:e453. URL: `http://dx.doi.org/10.7717/peerj.453`. doi:`10.7717/peerj.453`.

[45] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

[46] Vakkila J, Lotze MT. Inflammation and necrosis promote tumour growth. Nature Reviews Immunology 2004;4(8):641–8.

[47] Shuttleworth J, Todman A, Naguib R, Newman B, Bennett M. Colour texture analysis using co-occurrence matrices for classification of colon cancer images. In: IEEE Canadian Conference on Electrical and Computer Engineering; vol. 2. 2002, p. 1134–9. doi:10.1109/CCECE.2002.1013107.

[48] Doyle S, Rodriguez C, Madabhushi A, Tomaszeweski J, Feldman M. Detecting prostatic adenocarcinoma from digitized histology using a multiscale hierarchical classification approach. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE; 2006, p. 4759–62.

[49] Sertel O, Kong J, Shimada H, Catalyurek U, Saltz JH, Gurcan MN. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. Pattern recognition 2009;42(6):1093–103.

[50] Diamond J, Anderson NH, Bartels PH, Montironi R, Hamilton PW. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. Human Pathology 2004;35(9):1121–31.

[51] Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszeweski J. AUTOMATED GRADING OF PROSTATE CANCER USING ARCHITECTURAL AND TEXTURAL IMAGE FEATURES. 4th IEEE International Symposium on Biomedical Imaging 2007;:1284–7URL: http://dx.doi.org/10.1109/isbi.2007.357094. doi:10.1109/isbi.2007.357094.

[52] Bilgin CC, Bullough P, Plopper GE, Yener B. ECM-aware cell-graph mining for bone tissue modeling and classification. Data mining and knowledge discovery 2010;20(3):416–38.

[53] Sharma H, Alekseychuk A, Leskovsky P, Hellwich O, Anand R, Zerbe N, et al. Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. Diagnostic pathology 2012;7(1):134.

[54] Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on 1973;3(6):610–21.

[55] Daugman JG. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. Acoustics, Speech and Signal Processing, IEEE Transactions on 1988;36(7):1169–79.

[56] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern recognition 1996;29(1):51–9.

[57] Tkalčič M, Tasič JF. Colour spaces: perceptual, historical and applicational background. In: EUROCON 2003. Computer as a Tool. The IEEE Region 8; vol. 1. IEEE; 2003, p. 304–8.

[58] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review 2010;33(1-2):1–39.

[59] Chen L. Learning Ensembles of Convolutional Neural Networks 2016;URL: `theorycenter.cs.uchicago.edu`.

[60] Sharma H. Presentation: Deep convolutional neural networks for histological image analysis in gastric carcinoma whole slide images [Video file]. `http://patho-wsi.charite.de/pub/ECDP2016/data/vid/SY14-Sharma.mp4`; 2016.