Histopathology Cancer Detection

Seminar: Introduction To Deep Learning

Hasso Plattner Institut

Hagen Marin, Jan Carlo Schmid

Summer Semester 2023

Contents

1 Introduction

The use of Artificial Intelligence (AI) in the medical context has become increasingly concrete in recent years. Even though first use cases are still in an early phase of validation and implementation, valuable insights regarding large-scale deployment are emerging. One area in which these developments have been particularly noticeable is the use of neural networks in learning patterns directly from raw data to classify images. Beside specialties such as gastroenterology, dermatology and radiology, these classifiers showed promising results in the field of pathology [1]. The histopathological assessment of samples like potentially tumorous tissue involves fixation in formalin, cutting, embedding in paraffin, staining with hematoxylin and eosin (H&E) and subsequent visual characterization by a trained pathologist using a microscope. Pathologists investigate the presence, subtype and other histological characteristics of tumor tissue according to standardized criteria. In addition, for many tumor types, more quantitative analyses like the grade, a counting of mitotic cells and tumor-infiltrating lymphocytes, a quantification of tumor budding as well as many other types of analyses depending on the tumor entity are required. Many of these tasks are labor intensive and not perfectly reproducible by and between observers. Additionally, many histopathology-related tasks ask for enduring attention to detail, which is likely to decrease as a pathologist fatigues. Issues a computer wouldn't be affected by [2].

Recent major strides in the application of neural networks in pathological tasks were achieved through the digitazing of glass slides into whole-slide images (WSI) [3]. Throughout various cancer types algorithms were able to identify areas of interest. Furthermore, it was shown that some algorithms can successfully classify the origin of the primary tumor as well as detect structural variants and genetic alterations from standard histopathology slides with an accuracy comparable to molecular tests. Beside a possible time-sparing effect for pathologist, they hereby provide valuable information even for clinical experts [4, 5, 2]. Other studies indicated that based on neural networks survival predictions for patients with cancer might be made more accurately than by using conventional grading and histopathological subtyping [6, 7]. All these results underline the enormous potential of Deep Learning in the field of pathology and particularly the revolutional role of Convolutional Neural Networks in cancer detection through the classification and segmentation of histopathological images.

2 Dataset

The data set used in the present work originates from the CAMELYON16 challenge for breast cancer metastasis detection, a whole-slide image data set that was published in 2017 to offer a straight-forward binary image classification task for computational pathology. The histological images used for this purpose were acquired and digitized from so-called sentinel lymph nodes that were removed from women with breast cancer as part of an ordinary examination for metastases. They were collected in two different university centers in the Netherlands and were stained with H&E, respectively. Among the set of 270 women with breast cancer whole-slide images of 110 were found to have nodal metastases. Annotations have been performed pixel-wise [8].

The data was downloaded from the kaggle competition "Histopathology Cancer Detection" [1]. Within that competition, called as PatchCamelyon (PCam) benchmark, one was asked to create an algorithm to identify metastatic cancer in small image patches. The PCam dataset contains 327,680 patches extracted from Camelyon16 at a size of 96×96 pixels [9].

To reduce loading time during training extensively batches of 128 patches have been built and stored as pickle objects. Based on a Hold-out technique the so-obtained data set was further split into three subsets. 70 percent of the patches were used for training, 15 percent each for validation and testing. The distribution of tumorous and non-tumorous patches in each of these sets is provided in the Google Colab.

3 Architecture and Training

To date, Convolutional Neural Networks are the most used deep learning algorithms applied on pathology image analysis [3]. Other possible algorithms that are commonly applied in similar tasks include an additive integration of a Recurrent Neural Network as well as the application of Fully Convolutional Networks (FCN), Multiple Instance

[1] for further information: https://www.kaggle.com/competitions/histopathologic-cancer-detection/overview

Learning (MIL) and Generative Adversarial Networks (GAN) [2]. Regarding the strong evidence of benefits of CNN-applications, the fact that the tissue of patients in our dataset was collected only at a single timepoint and that the set was estimated to be fairly large, a CNN was chosen to perform the classification task.

Among common CNNs, several approaches have been developed and analysed in recent years. Studies that applied CNNs on similar data sets used VGG-Net-like architectures to classify different types of tissue [10] or to infer (molecular) subtypes [11, 12] from histologic images in cancer patients. More recent approaches performed digital pathology with the residual network architecture (ResNet), introduced by Microsoft in 2015. On its introduction, ResNet outperformed previous architectures by significant margins in all main tracks of the ImageNet computer vision competition, including object detection, classification, and localization [13]. Based on these results classification of benign vs. malignant cell tiles [14] or the detection of cancer gene mutations [4] have been conducted via pre-trained ResNet algorithms. Comparisons between different CNN-algorithms concluded that a pre-trained ResNet-algorithm performed impressively well on differentiating several subtypes of cancer [15].

3.1 Machine Learning Model

Regarding these results in current literature different ResNet algorithms were built for the task of pathological classification. First published in 2015, the ResNet architecture covers the issue of building deeper neural networks with more layers doesn't present the perfect solution for every new problem as accuracy rates tend to degradate, training error increases, and optimization gets harder. For our task two different ResNet models have been evaluated including a smaller model with 10 layers and a larger, pre-trained 18-layer model were built using Pytorch [2].

They were both trained under several approaches. The key concept of all ResNet models is the so-called residual block built by the identity function, a short-circuit convolution, and a regular block with two weighted layers and an activation function in between. In cases of different dimensions two solutions are possible: First, the shortcut still performs identity mapping, with extra zero entries padded for increasing dimensions. This option introduces no extra parameter. Second, the projection shortcut is used to match dimensions (done by 1×1 convolutions).

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 1: Architectures of different ResNet Models, derived from: He et al. [13]

As to be seen in *Figure 1* the ground architecture of a standard ResNet18 starts with a stacking of a convolutional layer [7x7], Batch Normalization and a ReLU-activation function which is followed by a max pooling layer [3x3]. Thereafter, in the ResNet18-model a structure of four different types of Layers containing each two repeated Basic Blocks is set. The blocks are stacked together one after the other. The Basic Blocks within the four layers only differ in the number of input and output channels. Each Basic Block consists of two Convolutional Layers. The first of these Convolutional Layers is followed by Batch Normalization and ReLU activation. The second is followed only by Batch Normalization. The final two layers are built by an Average Pooling layer and a fully connected layer resulting in $1.8x10^9$ FLOPs (floating point operations).

[2] for further information: https://debuggercafe.com/implementing-resnet18-in-pytorch-from-scratch/

3.2 Training

Training was conducted on one side with a hold-out technique after splitting the data with 70 percent shown during training, 15 percent in validation and 15 percent in during testing. On the other side a 4-fold cross validation was performed with 100 epoches including 1095 batches with 128 patches. The distribution of the dataset in the 4-fold approach was set to 63.75 percent on training, 21.25 percent on validation and 15 percent on testing, respectively.

3.3 Experiments

To conteract overfitting two different approaches have been implemented. First, as a regularization method to diversify weights and rely less on single features a dropout layer was applied after the third Basic Block and before the final fully connected layer with dropout probability of 0.6 and 0.5, respectively. The Dropout mechanism was introduced only in the ResNEt18-model, since the ResNet10 architecture was seen in training to be too small to benefit.

Second, Data Augmentation techniques (rotation, flips,...) during training and Test Time Augmentation (TTA) during validation and testing were used. The former was done by adding rotation between +/-180° and a chance of horizontal flip with 0.3. In the latter four slightly modified versions of the same image were included and the network was asked to predict the class of each of them during validation and testing. The probabilities for each of the modified versions are than summed up and averaged. By that technique, beside the correct answers one also averages the errors. The error can be big in a single vector, leading to a wrong answer, but when summed up the correct answer gets more robust. As pathological slides aren't changed in their meaning when largely rotated or flipped Test Time augmentation seems to be a valid technique to improve classification results and prevent overfitting.

3.4 Additional Hyperparameters

As additional hyperparameters a Adam optimizer with a standard learning rate of 0.001 and a weight decay of 0 a BCE-loss, an early stopping patience set to 10 and a min_delta set to 0.01 were implemented.

4 Evaluation

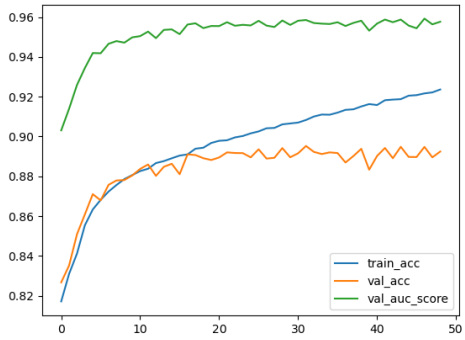
Based on our findings, the k-fold approach was not superior to the results obtained by the hold-out technique but came with enormous computational costs. Because of that reason it wasn't further investigated after validation. The focus of our work remained on the models using the prescribed hold-out technique.

4.1 Strategies and Metrics

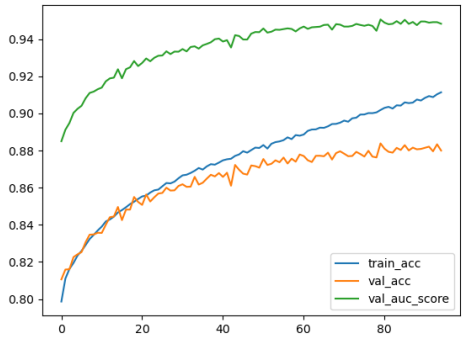
To evaluate the performance of the built model several statistical metrics which are widely used in current literature have been conducted. On the ResNet10 as well as the ResNet18-algorithm accuracy, recall, f1-score and the ROC-AUC were calculated with regard to Test Time augmentation but also without. Furthermore, the decision boundary was evaluated of whether it was accurate or should be adapted.

4.2 Model performance

As presented in *Figure 2* both models, ResNet10 and ResNet18, performed reasonable well on the validation set. As to be seen later in testing, the ResNet18 provided better results after data augmentation techniques have been put in place, while the ResNet10 model performed worse than without that additional increase in variety.



(a) ResNet10 without additional Data augmentation



(b) ResNet18 with two Dropout-Layers and TTA

Figure 2: Performance of ResNet models in Traning and Validation

As to be seen in the *Table 1* the ResNet10 algorithm performed better without an additional augmentation of the data. Nevertheless, both ResNet10-algorithms didn't reach the the performance of the ResNet18 model. Here, the data augmentation showed beneficial effects on all calculated statistical metrics. The evaluation of the decision boundary revealed it to be accurate as the maximum could be shown at 0.5. Further adaptation of the decision boundary was seen to be unnecessary.

	Resnet10 without TTA	ResNet10 with TTA	ResNet18 without TTA	ResNet18 with TTA
Accuracy	0.885	0.735	0.880	0.885
F1-score	0.854	0.657	0.851	0.857
AUC-score	0.951	0.815	0.948	0.951
Recall	0.837	0.632	0.854	0.857

Table 1: Performance of tested all ResNet models.

5 Discussion

Several aspects of our approach are worthy to be discussed. As said, the model that reached the best performance in our task was built out of 18 layers. In contrast, most scientific studies which developed similar tumor classifiers used models of a larger size. ResNet-architectures of 18, 34, 50, 101 and 152 layers were published. However, most of these studies focused on a classification of several different subtypes of cancer or compared different sizes of ResNet models against each other. Korbar et al. 2017 tested different variations of ResNet-architectures with 50–152 computational layers to identify five different types of colon polyps and normal tissue. Best performance was reached by the ResNet152 algorithm after 200 epochs with an overall accuracy of 93.0 % , Precision 89.7 %, Recall 88.3% and F1 Score 88.8%. [16]. A similar approach was chosen by Wei et al. 2020 to classify four major colorectal polyp types with a model that comprised five ResNet-architectures of 18, 34, 50, 101 and 152 layers, respectively [12]. Because the goal of our model was to differentiate between two classes, namely metastatic and non-tumorous tissue, a smaller model was considered sufficient.

The dataset our model was trained, validated and tested on derives from the PCam benchmark. Unlike most of the current literature the 327,680 patches were only of a size of 96x96 pixels. Kather et al. 2019 trained a ResNet18 model to detect GI cancer tissue with an out-of-sample area under the curve (AUC) of > 0.99. The used data set form the Cancer Genom Atlas consisted slides of 1554 patients split into more than 250,000 tiles with a size 224x224 pixels [4]. While that resolution is frequently used in pathological classification, other models were trained on data sets with even higher resolution[16, 17]. These observations suggest that the information our model could learn was inferior to the one of other data sets. Another point concerning the used dataset is given by the fact that the ID of the WSI each patch was derived from was not known and WSI weren't separated into training, validation and test set before splitting. Therefore, it seems probable that highly-correlated patches could be found in the different

sets. As it is not possible to overcome this issue any longer, it should at least be considered when evaluating models trained on the PCam benchmark.

Pathological data sets are commonly preprocessed. Oftentimes, tiles get color-normalized before presented to a model [4]. Since histological slides have no specific directions, data augmentation is regularly performed by random rotation, flipping and mirroring or color jittering[15, 12]. That regularly helps to increase variety of a data set. The here used models ResNet10 and ResNet18 were trained and tested with but also without additional data augmentation techniques. As described above, beside data augemntation by rotation and flipping the TTA method was implemented in our models. It brings up a similar idea in presenting slightly modified versions of the same image into the data set, sums predictions up and averages them to reduce error rates.

Our ResNet18 was pretrained on the ImageNet Database, which corresponds to the common procedure in similar scientific tasks [4, 15].

Based on our results, one could argue that a smaller ResNet model like the demonstrated ResNet10 shows reasonnably good results when resources are limited. Nevertheless, one should keep in mind that its statistical scores reveal a certain lack of robustness and thereby a lack of generalization, if the dataset differs from the given set in this task. While it performs rather well on the given testset, it is likely to overfit on the whole dataset as data augmentation doesn't improve the results but slightly worsens them. To cover this problem, new datasamples would probably require a certain transformation and normalisation to match the PatchCamelyon dataset. Other studies have overcome the potential issue of overfitting to the whole given dataset by conducting external evaluation.

Wei et al. 2020 classified four major colorectal polyp types with a model that comprised five ResNet-architectures of 18, 34, 50, 101 and 152 layers, respectively. 200 epochs of Training, the validation and an internal testing were performed on 7000 fixed-sized 224x224-pixel patches per class derived from 508 slides. Thereafter, the study further conducted an external validation with 238 slides from 24 different institutions in the US to examine its generalization abilities. Performance was measured against results of five local pathologists. On both testing sets the algorithm delivered human-level results with a decrease when performed on external data (Acc: int 93.5% vs. ext 87.0%, Recall: int 86.8% vs. ext 77.7% Spec: int 95.7% vs. ext 91.6%) indicating potential problems on more diverse data sets [12]. Similarly, Song et al. 2019 performed a ResNet34-model based on DeepLab V2 as well as a non-pretrained ResNet50 to classify colorectal carcinoma. Training was conducted on 113 090 adenomatous tiles (320x320 pixels) and 90 122 normal ones for training taken from 177 cases. On internal testing the ResNet34 model showed a AUC of 0.92 on 194 WSI cut into tiles of 2000x2000 pixels. On an external testing with slides from two other hospitals the DeepLab2-based model predicted 155 of 168 slides correctly [17].

As it counts for every new technique performed in a clinical setting, classification algorithms have to show that they can perform at least equal level if not better than the current state-of-the-art which is a human pathologist. So far, there is no publicly available study that tested pathologists on the PCam dataset. Therefore, a comparison of our model against human performance is not possible. Other studies have already shown promissing results of ResNet algorithms when tested against professional pathologists. For example, Wei et al. measured performance of their model on internal and external validation against results of five local pathologists in classifying four major colorectal polyp types. On both testing sets the algorithm delivered human-level results with a decrease when performed on external data (Acc: int 93.5 vs. ext 87.0, Recall: int 86.8 vs. ext 77.7 Spec: int 95.7 vs. ext 91.6) [12]. However, as the task as well as the built model differs from our own approach one has to be careful with comparisons.

As it counts for many of approaches that study the capabilities of neural networks in image classification our project faces the problems of a retrospective analysis for a single use case with limited pre-selected data. As shown in our ResNet10-model even with a dataset of the size of PCam, models tend to overfit and show low performance in prospective studies. So far only a limited number of studies have been published that underwent external validation, prospective evaluation and diverse metrics in order to investigate the full impact of Neural Networks in the clinical practice [1].

6 Conclusion

As already been said, there remain yet insufficiently answered challenges for a safe and effective deployment of AI in the complex ethical, technical and human-centered field of medicine. Nevertheless, algorithms which are used in pathology to distinguish malignant from physiological tissue might be among the first applications of AI that can be successfully implemented in a clinical routine setting. Furthermore, the development of new technologies is helping to continuously refine digital pathology approaches and promises to be able to extract a wealth of hidden information from routinely obtainable tissue slides. While the trained ResNet18 model with TTA showed promissing results, it faces the issue of a missing translation into medical practice. Further prospective studies on external datasets would be necessary to get essential information about its actual value.

References

[1] Pranav Rajpurkar et al. "AI in health and medicine". eng. In: *Nature Medicine* 28.1 (Jan. 2022), pp. 31–38. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0.

[2] Artem Shmatko et al. "Artificial intelligence in histopathology: enhancing cancer research and clinical oncology". en. In: *Nature Cancer* 3.9 (Sept. 2022), pp. 1026–1038. ISSN: 2662-1347. DOI: 10.1038/s43018-022-00436-4. URL: https://www.nature.com/articles/s43018-022-00436-4 (visited on 07/13/2023).

[3] Kaustav Bera et al. "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology". en. In: *Nature Reviews Clinical Oncology* 16.11 (Nov. 2019), pp. 703–715. ISSN: 1759-4774, 1759-4782. DOI: 10.1038/s41571-019-0252-y. URL: https://www.nature.com/articles/s41571-019-0252-y (visited on 07/13/2023).

[4] Jakob Nikolas Kather et al. "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer". en. In: *Nature Medicine* 25.7 (July 2019), pp. 1054–1056. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-019-0462-y. URL: https://www.nature.com/articles/s41591-019-0462-y (visited on 07/13/2023).

[5] Anne Laure Le Page et al. "Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology HES images". en. In: *Scientific Reports* 11.1 (Dec. 2021), p. 23912. ISSN: 2045-2322. DOI: 10.1038/s41598-021-03206-x. URL: https://www.nature.com/articles/s41598-021-03206-x (visited on 07/13/2023).

[6] Yu Fu et al. "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis". en. In: *Nature Cancer* 1.8 (July 2020), pp. 800–810. ISSN: 2662-1347. DOI: 10.1038/s43018-020-0085-8. URL: https://www.nature.com/articles/s43018-020-0085-8 (visited on 07/13/2023).

[7] Pierre Courtiol et al. "Deep learning-based classification of mesothelioma improves prediction of patient outcome". en. In: *Nature Medicine* 25.10 (Oct. 2019), pp. 1519–1525. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-019-0583-3. URL: https://www.nature.com/articles/s41591-019-0583-3 (visited on 07/13/2023).

[8] Babak Ehteshami Bejnordi et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". eng. In: *JAMA* 318.22 (Dec. 2017), pp. 2199–2210. ISSN: 1538-3598. DOI: 10.1001/jama.2017.14585.

[9] Bastiaan S. Veeling et al. *Rotation Equivariant CNNs for Digital Pathology*. arXiv:1806.03962 [cs, stat]. June 2018. URL: http://arxiv.org/abs/1806.03962 (visited on 07/13/2023).

[10] Babak Ehteshami Bejnordi et al. "Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies". en. In: *Modern Pathology* 31.10 (Jan. 2018), pp. 1502–1512. ISSN: 08933952. DOI: 10.1038/s41379-018-0073-z. URL: https://linkinghub.elsevier.com/retrieve/pii/S0893395222033014 (visited on 07/19/2023).

[11] Heather D. Couture et al. "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype". eng. In: *NPJ breast cancer* 4 (2018), p. 30. ISSN: 2374-4677. DOI: 10.1038/s41523-018-0079-1.

[12] Jason W. Wei et al. "Evaluation of a Deep Neural Network for Automated Classification of Colorectal Polyps on Histopathologic Slides". eng. In: *JAMA network open* 3.4 (Apr. 2020), e203398. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.3398.

[13] Kaiming He et al. *Deep Residual Learning for Image Recognition*. arXiv:1512.03385 [cs]. Dec. 2015. URL: http://arxiv.org/abs/1512.03385 (visited on 08/08/2023).

[14] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". en. In: *Nature Medicine* 25.8 (Aug. 2019), pp. 1301–1309. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-019-0508-1. URL: https://www.nature.com/articles/s41591-019-0508-1 (visited on 07/13/2023).

[15] A. Ben Hamida et al. "Deep learning for colon cancer histopathological images analysis". eng. In: *Computers in Biology and Medicine* 136 (Sept. 2021), p. 104730. ISSN: 1879-0534. DOI: 10.1016/j.compbiomed.2021.104730.

[16] Bruno Korbar et al. "Deep Learning for Classification of Colorectal Polyps on Whole-slide Images". eng. In: *Journal of Pathology Informatics* 8 (2017), p. 30. ISSN: 2229-5089. DOI: 10.4103/jpi.jpi_34_17.

[17] Zhigang Song et al. “Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists”. eng. In: *BMJ open* 10.9 (Sept. 2020), e036423. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2019-036423.