M. Bengs, S. Pant, M. Bockmayr, U. Schüller, and A. Schlaefer

Multi-Scale Input Strategies for Medulloblastoma Tumor Classification using Deep Transfer Learning

Abstract: Medulloblastoma (MB) is a primary central nervous system tumor and the most common malignant brain cancer among children. Neuropathologists perform microscopic inspection of histopathological tissue slides under a microscope to assess the severity of the tumor. This is a time-consuming task and often infused with observer variability. Recently, pre-trained convolutional neural networks (CNN) have shown promising results for MB subtype classification. Typically, high-resolution images are divided into smaller tiles for classification, while the size of the tiles has not been systematically evaluated. We study the impact of tile size and input strategy and classify the two major histopathological subtypes—Classic and Desmoplastic/Nodular. To this end, we use recently proposed EfficientNets and evaluate tiles with increasing size combined with various downsampling scales. Our results demonstrate using large input tiles pixels followed by intermediate downsampling and patch cropping significantly improves MB classification performance. Our top-performing method achieves the AUC-ROC value of 90.90% compared to 84.53% using the previous approach with smaller input tiles.

Keywords: Transfer learning, convolutional neural networks, digital pathology, histopathology, medulloblastoma

https://doi.org/10.1515/cdbme-2021-1014

M. Bengs, S. Pant, A. Schlaefer: Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany,
Email: marcel.bengs@tuhh.de
M. Bockmayr, U. Schüller: Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg 20246, Germany
M. Bockmayr, U. Schüller: Research Institute Children's Cancer Center Hamburg, Martinistraße 52, Hamburg 20251, Germany
U. Schüller: Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg 20246, Germany
M. Bockmayr: Mildred Scheel Cancer Career Center HaTriCS4, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany

1 Introduction

Medulloblastoma (MB) is the most common malignant brain tumor in children and a major cause of morbidity, as well as mortality in pediatric oncology [15]. All MBs are classified as Grade IV tumors by the World Health Organization (WHO) [10], indicating they are invasive and fast-growing. The 2016 edition of the World Health Organization Classification of Tumors of the Central Nervous System (CNS) has defined four histological subtypes of MB [10, 12]— classic type (CMB), desmoplastic/nodular type (DN), MB with extensive nodularity (MBEN), and large cell anaplastic MB (LCA). Each subtype is associated with different prognoses and therapies, while early and precise diagnosis is vital for increasing the survival rates for patients [13].

For establishing a diagnosis, a tissue specimen or biopsy sample is extracted from the suspected region of the brain. Then, neuropathologists assess the tissue slides under the microscope or digitize the magnified view to obtain an extremely high-resolution image which is also called Whole Slide Image (WSI). To detect and discern different types and stages of tumors, they apply human-based decision rules based on their skills, experience, and knowledge to detect and discern different types and stages of tumors. However, the visual assessment of such tissue scans is a laborious and time-consuming task, which is also affected by inter-observer variability [2]. These problems have emphasized the requirement of automated decision support tool [17].

One way to implement automated classification of MB subtypes is by means of manual feature extraction [5]. Although this approach allows for promising results, manual feature extraction is task-dependent and requires strong domain expertise. In contrast to that, Convolutional Neural Networks (CNNs) provide a more general approach and it has been demonstrated recently that CNNs outperform conventional methods in various pathological image analysis tasks [1,16]. While CNNs provide a general approach with superior performance in many learning tasks, they require a large number of training examples. However, in rare cancer

such as MB, there are not enough training data available to train any powerful CNN architecture. To counter this problem, transfer learning is typically used in digital pathology [9, 14].

To mitigate the problem of scant training examples in MB classification, a recent study has compared various benchmark CNN architectures together with transfer learning [3]. The study demonstrates that EfficientNets [18] with a larger input resolution outperform classical CNN architectures with smaller input resolutions. However, pre-trained CNNs are optimized for a fixed input resolution, e.g., 224 × 224 pixels [18], which conflicts with the high-resolution WSIs. Hence, WSIs are typically divided into several thousand tiles [1, 7], which are processed with a deep learning approach afterwards. Here, the question arises which tile size to choose.

We systematically study the effect of tile size, image downsampling, and input strategy for the task of MB subtype classification using pre-trained CNNs. We use a data set with WSI from 161 different patients and consider the task of differentiating between types CMB and DN.

2 Materials and Methods

2.1 Data Set

We use a dataset collected from 12 clinical sites in Germany from 1989-2011. All local institutional guidelines were followed including informed consent from the patients. Slides were stained by hematoxylin and eosin (H & E) and scanned at the same institution with a magnification of 200x. Neuropathologists have labeled the images as Classic (CMB) or Desmoplastic/Nodular (DN). The data contains WSIs of 161 patients of which 103 cases are CMB and 58 are DN. Each WSI has more than one cancerous region. To generate a data set consisting of image tiles, neuropathologists examined the WSIs and identified representative cancerous regions. Afterwards, we extracted tiles with a size of 2000 × 2000 pixels from the cancerous regions. Each patient contains multiple labeled tiles. There are 1574 tiles for 103 patients with CMB and 1195 tiles for 58 patients with DN cases.

We evaluate three different extracted tile sizes $(h_t \times w_t)$. Given an extracted tile with a size of 2000 × 2000, we crop larger tiles with a size of 4000 × 4000 pixels and 8000 × 8000 pixels such that the manually extracted tile is centered. In this way, all sets share the same center area, while the large ones also include more overall context. We evaluate our models based on tile classification performance.

2.2 Deep Learning Methods

We follow the concept of a previous study on MB classification [3] and consider pre-trained EfficientNets [18]. The key advantage of EfficientNets is the compound scaling method, which uniformly scales network width, depth, and input resolution starting with the baseline EfficientNet-B0. Considering the findings of previous work [3], we focus on EfficientNet-B0 (E#Net-B0) with an input resolution of 224 × 224 and EfficientNet-B5 (E#Net-B5) with an input resolution of 456 × 456. Note, we use architectures pre-trained on ImageNet.

Next, we study the relation between tile size, input strategy, and the corresponding classification performance. Our general classification pipeline is shown in Figure 1.

Given the sets with different tile sizes, we first follow the previous approach [3] and simply downsample an entire image tile to the corresponding input resolution $(h_i \times w_i)$ of the CNN. This leads to extreme downsampling of the larger tiles, especially for 8000 × 8000. Hence, we also consider an additional input strategy, where we first downsample the image tile to an inter-mediate resolution $(\hat{h}_t \times \hat{w}_t)$ larger than the CNN input. Then, during training, we randomly crop input patches $(h_i \times w_i)$ from the intermediate tiles, and during evaluation we take multiple ordered crops and average the predictions for an intermediate tile afterwards. We consider three different square intermediate tile sizes $\hat{h}_t, \hat{w}_t \in \{456, 1000, 2000\}$. Note, when we combine the extracted tiles with a size of 2000 × 2000 with the intermediate tile size of 2000 × 2000 no downsampling is performed. We randomly split our data based on patients and consider 10-fold cross-validation. In each fold, data is divided into a training set comprising 139 patients, and a test and validation set comprising of 7 patients each. The test and validation subsets all consist of five and two cases for type CMB and DN, respectively. To counter class imbalance during training, we weight the loss of the individual classes inversely proportional to samples of each class. We employ data augmentation during training using brightness, contrast, saturation, and hue augmentation as well as random horizontal and vertical flipping of the images. The training is implemented with 300 epochs for all 10-folds with a batch size of 15.

3 Results

We report the area under the receiver operating curve (AUC) with 95% confidence intervals (CI) using bias-corrected and accelerated bootstrapping with n_{CI} = 10,000 bootstrap samples in Table 1. For testing of significance, we use a permutation

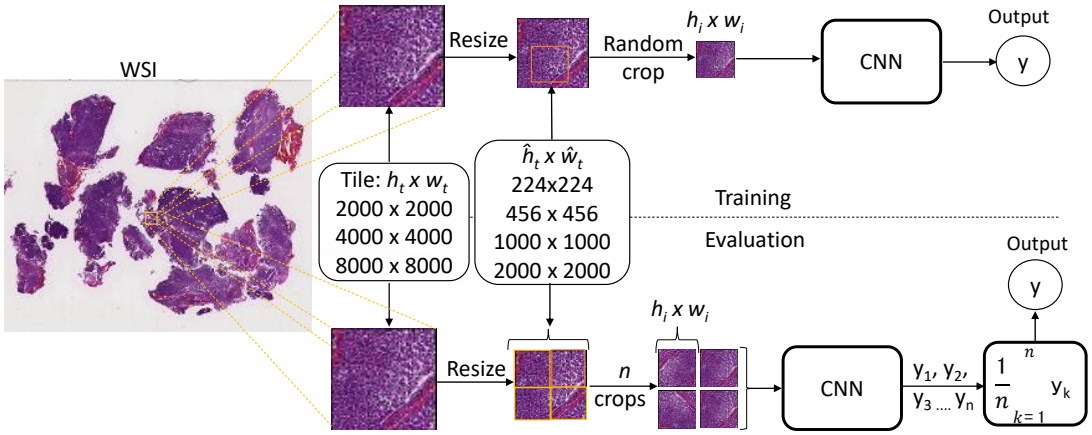

Figure 1: Our classification pipeline for the histological subtypes classic (CMB) and desmoplastic/nodular (DN). A tile ($h_t \times w_t$) is extracted from the WSI and is then downsampled to a size of $\hat{h}_t \times \hat{w}_t$. Afterwards, we take a randomly localized crop with the size of the network's input resolution $h_i \times w_i$ during training. For evaluation, we take n ordered crops and average all crop prediction to obtain one final prediction y.

Table 1: Results for all experiments given in percent. The best performing method is shown in bold. 95% CIs are provided in brackets. Note, the input resolution of E#Net-B0 and E#Net-B5 are 224 × 224 px and 456 × 456 px, respectively.

Tile	$\hat{h}_t \times \hat{w}_t$	E#Net-B0 AUC	E#Net-B5 AUC
2000	224 × 224	80.24(77 – 83)**[3]**	–
×	456 × 456	84.91(82 − 87)	84.53(81 − 87)**[3]**
	1000 × 1000	84.29(82 − 87)	86.93(84 − 89)
2000	2000 × 2000	79.92(77 − 83)	81.73(79 − 84)
4000	224 × 224	82.67(81 − 86)	–
×	456 × 456	82.97(79 − 85)	85.27(82 − 88)
	1000 × 1000	84.03(82 − 87)	89.63(87 − 91)
4000	**2000 × 2000**	86.42(84 − 88)	**90.90(89 − 93)**
8000	224 × 224	77.72(74 − 80)	–
×	456 × 456	82.95(79 − 85)	83.74(81 − 86)
	1000 × 1000	84.03(82 − 86)	88.86(87 − 91)
8000	2000 × 2000	84.59(81 − 86)	90.15(88 − 92)

test with n_P = 10,000 samples and a significance level of α = 5% [6]. Our results show that E#Net-B5 outperforms E#Net-B0 for all our experiments, except for an intermediate tile resolution of 456 × 456. Also, our results demonstrate that using a tile with size of 4000 × 4000 pixels, downsampled to an intermediate size of 2000 × 2000 pixels works best and significantly ($p < 0.05$) outperforms the previous approach [3] that used a smaller tile size of 2000 × 2000 px downsampled to 456 × 456 px.

4 Discussion

We consider MB subtype classification using pre-trained EfficientNets and study the impact of input patches with different scales and global context for this task. Our results in Table 1 demonstrate that using the previous approach [3] with larger tiles downsampled to the network input resolution only led to minor performance improvements for a tile size of 4000 × 4000 px. However, for extremely large tiles (8000 × 8000 px) performance is even reduced. This indicates that using larger tiles is beneficial, however, when too much downsampling is performed relevant feature are lost. Similar, when no downsampling is performed performance is reduced, i.e., in the case of a tile size and an intermediate resolution of 2000 × 2000 px. This indicates that here the global context is missing, while high resolution information is preserved. This demands a method to preserve global context without sacrificing the fine-grained details. Our results highlight that taking larger tiles, followed by intermediate downsampling and multi-cropping during training enables the right trade-off between the exploitation of global context and the preservation of detailed information. Our results demonstrate that this significantly improves the classification performance. Also, this superior performance might be linked to a simple version of multiple-instance learning (MIL) [4]; the predictions are averaged in our study which acts as a pooling function in MIL terminologies. So far, we only consider MIL during evaluation, and considering more advanced versions of MIL during training like attention-based MIL [8] could lead to

promising results. Also, WSI classification remains an open challenge and could be addressed by combining our findings with recent works on WSI classification [7, 11].

5 Conclusion

We address the task of MB tumor classification and study the impact of input patches with different scales and global contexts. Our results highlight that including more overall image context is beneficial. However, simply downsampling larger tiles that cover enlarged image areas directly to the input resolution of a CNN does not lead to any performance improvement. Instead, downsampling to an intermediate resolution followed by a multi-cropping strategy significantly boosts performance. Future work could focus on evaluating more advanced MIL techniques and on classifying all subtypes of MB using a larger data set.

Acknowledgment: This work was partially supported by the Hamburg University of Technology i3 initiative.

References

[1] M. Z. Alom, T. Aspiras, T. M. Taha, V. K. Asari, T. Bowen, D. Billiter, and S. Arkell. Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases. arXiv preprint arXiv:1904.09075, 2019.

[2] J. Arevalo, A. Cruz-Roa, et al. Histopathology image representation for automatic analysis: A state-of-the-art review. Revista Med, 22(2):79–91, 2014.

[3] M. Bengs, M. Bockmayr, U. Schüller, and A. Schlaefer. Medulloblastoma Tumor Classification using Deep Transfer Learning with Multi-Scale EfficientNets. In Medical Imaging 2021: Digital Pathology, volume 11603, page 116030D. International Society for Optics and Photonics, 2021.

[4] H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer. Multiple instance learning for heterogeneous images: Training a cnn for histopathology. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 254–262. Springer, 2018

[5] D. Das, L. B. Mahanta, S. Ahmed, and B. K. Baishya. Classification of childhood medulloblastoma into who-defined multiple subtypes based on textural analysis. Journal of microscopy, 279 (1):26–38, 2020.

[6] B. Efron and R. J. Tibshirani. An introduction to the bootstrap. CRC press, 1994.

[7] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports, 10(1):1–11, 2020.

[8] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In International conference on machine learning, pages 2127–2136. PMLR, 2018.

[9] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE, 2017.

[10] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta neuropathologica, 131(6):803–820, 2016.

[11] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, pages 1–16, 2021.

[12] P. A. Northcott, G. W. Robinson, C. P. Kratz, D. J. Mabbott, S. L. Pomeroy, S. C. Clifford, S. Rutkowski, D. W. Ellison, D. Malkin, M. D. Taylor, et al. Medulloblastoma. Nature Reviews Disease Primers, 5(1):1–20, 2019.

[13] P. A. Northcott, G. W. Robinson, C. P. Kratz, D. J. Mabbott, S. L. Pomeroy, S. C. Clifford, S. Rutkowski, D. W. Ellison, D. Malkin, M. D. Taylor, et al. Medulloblastoma. Nature Reviews Disease Primers, 5(1):1–20, 2019.

[14] H. T. H. Phan, A. Kumar, J. Kim, and D. Feng. Transfer learning of a convolutional neural network for hep-2 cell image classification. In 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pages 1208–1211. IEEE, 2016.

[15] I. F. Pollack and R. I. Jakacki. Childhood brain tumors: epidemiology, current management and future directions. Nature Reviews Neurology, 7(9):495–506, Sep 2011. ISSN 1759-4758, 1759-4766. 10.1038/nrneurol.2011.110.

[16] V. Rachapudi and G. L. Devi. Improved convolutional neural network based histopathological image classification. Evolutionary Intelligence, pages 1–7, 2020.

[17] A. Stenzinger, M. Alber, M. Allg uer, P. Jurmeister, M. Bockmayr,J. Budczies, J. Lennerz, J. Eschrich, D. Kazdal, P. Schirmacher,et al. Artificial intelligence and pathology: from principles to practice and future applications in histomorphology and molecular profiling. In Seminars in cancer biology. Elsevier, 2021.

[18] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pages 6105–6114. PMLR, 2019.