Predicting the 1p/19q co-deletion status of low-grade glioma from MR images using Convolutional Neural Network

Author: Anja Lieberherr
Supervised by: Dr. Sarah Brüningk
Labrotation
11.10.2021 - 14.11.2021
ETH Zurich
D-BSSE

Contents

1 Introduction **3**

2 Methodology **5**

2.1 Included data 5

2.1.1 Tumor segmentation 5

2.2 CNN model 6

2.2.1 Implementation 7

2.2.2 Performance metrics 7

2.3 Hyper-parameter tuning 8

3 Results **9**

3.1 Overall Performance 9

3.1.1 Performance for 1000 Patches per Patient 10

3.1.2 Performance for 60 Patches per Patient 12

4 Discussion and Outlook **13**

5 Appendix **15**

6 Bibliography **17**

Abstract

Patients having Low-grade gliomas (LGG) with the 1p/19q co-deletion mutation have been shown to better respond to treatment than patients without the mutation translating to improved survival prognosis, too. However, current methods to detect this mutation are brain-tissue biopsy or surgical resection of the tumor which are highly invasive. The aim of this lab rotation is to use a simple convolutional neural network (CNN) to predict the 1p/19q status from magnetic resonance (MR) images as a non-invasive method. The data comprised a total of 112 LGG patients with biopsy-proven 1p/19q status. CNN hyperparameters were optimized by grid search leading to a 5-fold cross-validation best mean(standard deviation) test set performance of 67.2%(7.3%) accuracy, 71.7%(7.6%) AUCroc and 76.3%(5.8%) average precision score. Hence, the model is not yet applicable for clinical use but may require further investigation of hyper-parameter tuning, inclusion of different contrast images as well as data augmentation to boost performance. Once the presented method is improved it could potentially be used as an alternative to surgical biopsy for predicting 1p/19q co-deletion status.

1 Introduction

Motivation

Amongst tumors from the central nervous system negatively affecting millions of patients worldwide, brain cancers account for the highest prevalence of more than ninety percent, Patel et al. (2019). Gliomas are the most frequent primary brain tumors originating in the brain, Cha (2006). The World Health Organisation (WHO) classifies them into four grades based on their aggressiveness, Kleihues et al. (1993). The grading scale of WHO is based on histological features and classifies tumors into levels of differing malignancy. Malignancy describes the tendency for a tumour to undergo anaplastic transformation over time, thus providing an essentially moving target. Low-grade gliomas (LGG) originate either from oligodendrocytes, astrocytes or oligoastrocytes and thus determine the type of the LGG, Network (2015), van den Bent et al. (2013). According to WHO, LGG are graded as either 2 or 3, Louis et al. (2007), Kleihues et al. (1993). The designation grade 2 is reserved for lesions with histologic evidence of malignancy, generally in the form of mitotic activity including clearly expressed infiltrative capabilities. Some tumour types tend to progress to lesions with higher grades of malignancy, namely to high-grade gliomas (HGG, WHO grade 4). LGG are less aggressive tumors with a better prognosis. The median survival time of LGG patients is highly variable, from a few months to more than 15 years, mainly depending on the clinical factors and molecular characteristics of the tumor, van den Bent et al. (2013), Macdonald et al. (1990), Franceschi et al. (2018). Presently, treatment includes active surveillance, surgery, radiotherapy, and chemotherapy either separately or in combination, Network (2015). Although histological grading of tumors is the standard procedure for diagnosis and subsequent treatment planning, it is known that histopathological diagnosis lacks information about other tumor properties (e.g., genomic biomarkers) that can impact optimal therapy options. Molecular biomarker analysis of LGG have shown that patients diagnosed with 1p/19q co-deletion mutation status had a significantly longer progression-free survival time and were more sensitive to therapeutics in terms of chemotherapy and radiotherapy compared to those with 1p/19q non-deleted tumors, Franceschi et al. (2018), Garcia et al. (2018), Ricard et al. (2007), van den Bent et al. (2013). Thus, the early detection of this chromosomal abnormality among patients diagnosed with LGG is of utmost importance. Currently, 1p/19q co-deletion is assessed from a tumor's histopathological sample through fluorescence in-situ hybridization (FISH) which is a time-consuming procedure, Scheie et al. (2006). The samples are collected via the brain-tissue biopsy or the surgical resection of the tumor, both highly invasive procedures, Quang-Hien Kha (2021). Identifying non-invasive methods to accurately predict the co-deletion status using image analysis with deep learning addresses the limitations of mentioned current procedure.

Related Work

Several studies have previously investigated whether the 1p/19q status can be predicted from medical images. Fellah et al. (2013) presented univariate analysis and multivariate random forest models to determine 1p/19q status of 50 LGG patients from multimodal magnetic resonance (MR) images including conventional MR images, diffusion-weighted imaging (DWI), perfusion-weighted imaging (PWI), and MR spectroscopy. DWI, PWI, and MRI spectroscopy showed no significant difference between tumors with and without 1p/19q loss in their study. The classification accuracy, to separate 1p/19q intact from 1p/19q co-deletion genotype between multimodal MR imaging and conventional MR images, was 40% misclassification error, and 79% sensitivity for multimodal MR imaging vs. 48% misclassification error, and 70% sensitivity for conventional MR images. They concluded that inclusion of DWI, PWI, and MR spectroscopy was not useful for determining 1p/19q status compared with conventional MR images. Jansen et al. (2012) presented detection of 1p/19q status from [^{18}F] fluoroethyltyrosine-PET (FET-PET) images. They derived several biomarkers from PET images and correlated these with 1p/19q status. The FET-PET-based predictions scored a sensitivity of 62% and a specificity of 85% with a total of 144 patients (BrainLab, Germany) included. They concluded that these biomarkers do not reliably predict the status of 1p/19q in individual patients. Iwadate et al. (2016) studied detection of 1p/19q co-deletion from ^{11}C-methionine PET images of 144 patients. They performed a separate analysis according to the WHO histological grade. A sensitivity of 88% and specificity of 64% in grade 2 tumours, and a sensitivity of 83% and specificity of 64% in grade 3 tumours were reported. The limitation of this study were the retrospective nature of the study design. DeAngelis (2001) describes MR imaging as a non-invasive medical imaging technique that provides outstanding soft-tissue contrast that has become the standard imaging technique for brain tumor diagnosis.

Akkus et al. (2017) predicted the co-deletion status from post-contrast T1 and T2-weighted MR images of 159 LGG patients using a multi-scale CNN approach. The data (477 slices) was downloaded from the brain tumor patient database at Mayo Clinic and was divided into training and test sets. A total of 90 slices were randomly selected from the data at the beginning, as a test set, and were never seen by the CNN during training. Twenty percent of the training data was separated as validation set during the training. 252 slices of the training set were randomly selected at each epoch for training. A 30-fold data augmentation was applied to the training set resulting in 7560 slices. Their model was overfitting to the training data when data augmentation was not used. The accuracy of the trained CNN without data augmentation remained below 80% for the test data. When data augmentation was applied to the training set at each epoch, to increase the training samples and to achieve generalization ability, their model achieved an accuracy of 87.7% on the test set. Quang-Hien Kha (2021) performed a similar study as Akkus et al. (2017), including a comparative performance analysis among different machine learning algorithms. The model XGBoost ranked first in accuracy with 69.2% and a significant performance in terms of the area under the receiver operating characteristic curve (AUCroc) and in terms of the area under the precision recall curve (AUPRC) (0.71 and 0.83, respectively). They also used a feature selection based on Shapley Additive Explanations (SHAP) values, Lundberg and Lee (2017).

The aim in this lab rotation is the prediction of 1p/19q status from T2-Fast-Spin-Echo (FSE) MR images, preprocessed into two-dimensional (2D) patches using a simple convolutional neural network (CNN). In contrast to previous approaches 2D images were used instead of three-dimensional (3D) images due to computational reasons and the limited number of samples.

2 Methodology

2.1 Included data

The Cancer Imaging Archive (TCIA) database of 159 LGG patients was used as for this study. All MR images were downloaded via the National Biomedical Imaging Archive (NBIA) Data Retriever, Can. Statuses of 1p/19q chromosomal arm were fully reported based on the standard histopathology examination via brain biopsy or surgical tumor resection. A total of 112 LGG patients showing the contrast type T2-Fast-Spin-Echo (FSE) were used as input to the classification algorithm, Sankowski et al. (2012). The dataset as a whole contained 60.7% double-deleted 1p/19q cases and 39.3% non-deleted 1p/19q cases. Furthermore, there were in total 66% cases as grade 2 classified LGGs and 34% cases as grade 3 classified LGGs. There was further variation with respect to tumor types: 11% patients had Astrocytoma, 28% patients comprised Oligodendroglioma, and a majority of 61% patients were attributed to Oligoastrocytoma. The data was split to perform a 5-fold nested cross-validation into training, validation and test sets. First, the dataset was split into 80% training data and 20% test data. The training data was further split to 75% training data and 25% validation data.

All the sets were balanced for 1p/19q status, grade, and tumor type using the sklearn "Stratified K-Folds cross-validator" function, skl. The analysis of the distribution of the tumor type, label and grade in the training, validation, and test sets revealed that the prevalence of all sets was comparable but the subtypes were represented in different portions. Especially, the subtype Astrocytoma is present in clear minority, see tab. 6, 7, 8.

2.1.1 Tumor segmentation

In addition to T2-FSE images, the Cancer Imaging Archive (TCIA) also provided segmentation masks for the included patients. The T2-FSE images were resized to the same dimension (256x256) as the segmentation images, whereas the third dimension varied depending on the volume of the tumor. The masks were then used to differentiate the tumor pixels from non-tumor pixels and to further localize and visualize the tumor, see fig. 1.

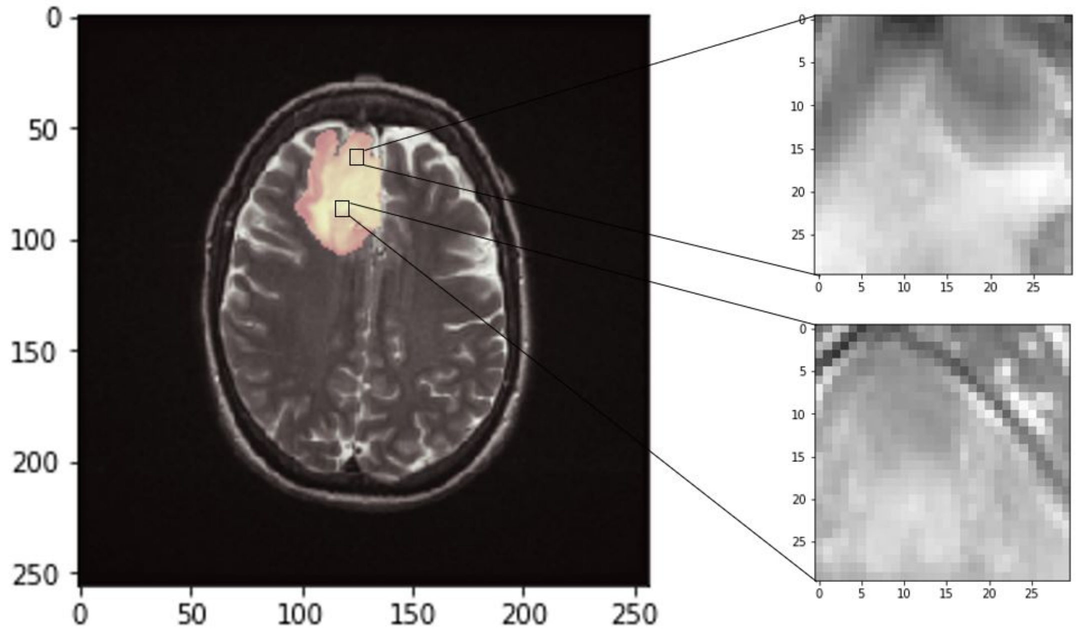


Figure 1: Represented example of a T2-FSE MR image with visualized tumor using the segmentation slice. Illustration of how the tumor was segmented and cut into the relevant image patches (miniatures).

After data exploration and tumor visualisations, the data was processed into 2D patches. Due to computational reasons, 2D tumor images instead of 3D images were used for this preliminary study. Only the segmentation slices containing at least a minimum number of tumor data were considered (minimum 40 tumor pixel per (256x256) segmentation slice). A binary boundary box on each segmentation slice, passing the threshold of tumor data, was defined to cut out the tumor data. Afterwards, patches of size (30x30) with a stepsize of 1 were produced in the area of the defined binary boundary box using the library 'patchify', pat. The size of the patches was defined to be large enough for tumor patterns to be recognized and small enough with respect to the microstructures in the smaller tumors. It was defined that only patches containing at least 50% tumor data were considered as input data to the classification algorithm. Furthermore, each patch was normalized to unit variance and zero mean using the sklearn "StandardScaler" function, skl. Depending on the size and shape of the tumor, the preprocessing resulted in a various number of patches per patient 9595(60, 116854) (median with full ranges). A certain number of patches per patient was defined to be served as input to the CNN. These patches were selected randomly for each patient. If a patient had less patches than the number of patches defined as input, all the patches available form that patient were included. Fig. 1 illustrates how the tumor was segmented and cut into the relevant image patches.

2.2 CNN model

In deep learning, a convolutional neural network (CNN) is a class of artificial neural network, mostly applied for computer vision tasks. A CNN consists of multiple layers, up to hundreds of layers for the very deep neural networks. A CNN typically has three different layers: a convolutional layer, a pooling layer, and a fully connected layer. The convolution layer is the core building block of the CNN. It carries the main portion of the network's computational load, Aghdam and Heravi (2017). Conventionally, the first convolutional layer captures the low-level features such as edges, color, gradient orientation, etc. With added layers, the architecture adapts to the high-level features as well, Korolev et al. (2017), Albawi et al. (2017). Similarly to the convolutional layer, the pooling layer is responsible for reducing the spatial size of the convolved features. This decreases the computational power required to process the data by reducing dimensionality. Furthermore, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training the model, Guo et al. (2017), Albawi et al. (2017). There are two types of pooling: max pooling and average pooling. Max pooling returns the maximum value from the portion of the image covered by the pooling kernel. On the other hand, average pooling returns the average of all the values from the portion of the image covered by the kernel, Lin et al. (2017). After going through the above process, the model is able to learn the features. Finally, the final output is flattened and fed to the last layer for classification purposes which is a fully connected layer, Albawi et al. (2017).

For the classification purpose in this study, a low number of convolutional layers with ReLu activation function were used, see fig. 2. Global average pooling was applied in the final pooling layer. The sigmoid activation function was applied in the output layer. The binary cross-entropy loss is used for binary classification applications. It computes the cross-entropy loss between true labels and predicted labels. The chosen optimizer Adam is a stochastic gradient descent optimization method that is based on adaptive estimation of first-order and second-order moments, Ker. According to Kingma and Ba (2014), the method is computationally efficient, has little memory requirement, is invariant to diagonal re-scaling of gradients, and is well suited for problems that are large in terms of data and parameters.

When using deep networks, in particular CNNs, there is a high risk of overfitting. This is a direct result of the large number of network parameters relative to the number of features provided by the MR images. The number of features available from MR images may not be sufficient to provide adequate learning and generalizability to the parameters in the network. There exist several methods to reduce overfitting in neural networks. The following three measures were applied in this study: early stopping, dropout and batch normalization. The principle of early stopping addresses the major challenge in training neural networks, e.g. how long to train them. Too little training will mean that the model will underfit the training and the test sets. Too much training will mean that the model will overfit the training dataset and have poor performance on the test set. A compromise is to train on the training dataset but to stop training at the point when performance on a validation dataset starts to degrade. Furthermore, dropout was applied after the MaxPooling layer to prevent overfitting by randomly dropping neurons from the neural network during training.

The effect of dropout is that the network becomes less sensitive to the specific weights of individual neurons and generalizes more, Srivastava et al. (2014). Batch normalization can also be used as a regularization technique. By adding batch normalization the internal covariate shift are reduced and instability in distributions of layer activations in deeper networks can reduce the effect of overfitting, Li et al. (2021).

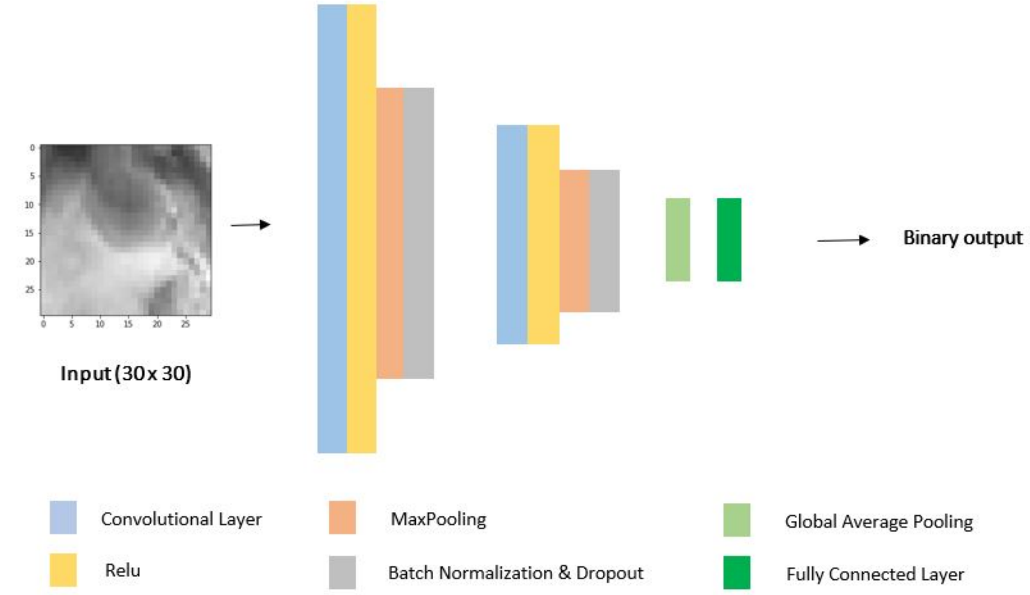


Figure 2: Architecture of used CNN showing how the binary output is generated. The input image is processed along convolutional layers, activation functions, regularizations, pooling layers and fully connected layers.

2.2.1 Implementation

All implementations are based on TensorFlow with Keras, Ker. A Datagenerator was used as a memory effective means of data loading. Instead of loading the entire dataset at once the class DataGenerator does the real-time data feeding to the model, dat.

2.2.2 Performance metrics

Recall, Precision, Accuracy, AUCroc and APS were used to measure the classification performance of the CNN and were computed as follows:

- TP = true positive
- TN = true negative
- FP = false positive
- FN = false negative
- ROC curve = receiver operating characteristic curve
- recall = TP / (TP + FN)
- precision = TP / (TP + FP)
- accuracy = (TP + TN) / (TP + TN + FP + FN)
- AUCroc = area under the ROC curve
- AUPRC = area under Precision-Recall curve
- APS = average precision score

2.3 Hyper-parameter tuning

The model was trained for 3 runs and 5 folds using a patience of 50 and 1000 epochs. In order to produce consistent results across different runs, the random seed in the random selection of patches per patient was set on the run index. The classification performance of the CNN architecture on the validation set was used to tune the hyper-parameters of the model. The accuracy, the area under the ROC curve as well as the average precision score were considered to select the best set of hyper-parameters. The learning rate is a key parameter to set as this parameter scales the magnitude of the weight updates in order to minimize the loss function. If the learning rate is set too low, training is not only slower but may become permanently stuck in a local minimum with a high training error. However, if the learning rate is set too high, it can cause undesirable divergent behavior in the loss function. The following values for different hyper-parameters were tested by Grid search.

- learning rate: 5e-3, 1e-3, 5e-4, 1e-4, 1e-5
- batch size: 50, 100
- number of layers: 2, 3
- kernel size: 3, 4, 5
- dropout: 0.25, 0.5
- batch normalization: 0, 1, 2

3 Results

Hyperparameter screen
The evaluation of the hyper-parameter screen revealed the best combination of hyper-parameters: Learning rate = 1e-5 , kernel size = 4, batch normalization = 2, dropout = 0.25, and number of layers = 2. The learning rate of 1e-5 showed the best performance compared to larger values. A kernel size of 4 revealed slightly better results (APS = 0.80) than a kernel size of 3 (APS = 0.78). By adding further layers to the network the risk to overfit does increase as long as the model performance is not stable. By comparing the APS of two layers (0.73) and the APS of three layers (0.79), one can conclude that two layers performed better than three layers. Increasing the batch size from 50 (APS = 0.79) to 100 (APS = 0.80) did not improve the model performance. Dropout values of 0.25 and 0.5 showed similar results while the performance using 0.25 (APS = 0.80) was slightly better than using 0.5 (APS = 0.76). Dropping units less often gives the units more opportunities to conspire with each other to fit the training set, Goodfellow et al. (2016).

3.1 Overall Performance

The performance varied depending on the number of patches per patient defined as input. To find out how the model performed on the different types of tumor, the accuracy per tumor subtype was calculated.

	Oligoastrocytoma	**Oligodendroglioma**	**Astrocytoma**
Prevalence	61.6%	27.7%	10.7%
Acc	71.7% ±5.8%	68.2% ±6.2%	55.0% ±10.0%

Table 1: Prevalence score of each tumor subtype and accuracy averaged over 3 runs and 5 folds according to tumor type in the training set. The values refer to the mean values of accuracy including standard deviation (std).

	Oligoastrocytoma	**Oligodendroglioma**	**Astrocytoma**
Prevalence	61.6%	27.7%	10.7%
Acc	71.4% ±9.5%	67.1% ±5.9%	49.6% ±14.2%

Table 2: Prevalence score of each tumor subtype and accuracy averaged over 3 runs and 5 folds according to tumor type in the validation set. The values refer to the mean values of accuracy including std.

	Oligoastrocytoma	**Oligodendroglioma**	**Astrocytoma**
Prevalence	61.6%	27.7%	10.7%
Acc	66.1% ±8.4%	62.5% ±6.8%	53.7% ±15.2%

Table 3: Prevalence score of each tumor subtype and accuracy averaged over 3 runs and 5 folds according to tumor type in the test set. The values refer to the mean values of accuracy including std.

The training and validation set are performing similar in terms of accuracy, indicating no overfitting, see tab. 1, 2. The performance on the test set for type Oligoastrocytoma (66.1%, 8.4%) (mean accuracy, std) is slighlty lower than on the validation set (71.4%, 9.5%) and on the training set (71.7%, 5.8%). Due to the low prevalence of type Astrocytoma and the fairly large uncertainties in general, the difference in performance on the test set (53.7%, 15.2%), on the training set (55.0%, 10.0%), and on the validation set (49.6%, 14.2%) is not significant, see tab. 1, 2, 3. The type Astrocytoma was not well represented in the training cohort, see tab. 6, and is the weakest performing subtype, see tab. 1, 2, 3. The subtype Oligoastrocytoma is the mostly represented type in the dataset, see tab. 6, 7, 8, and shows also the highest accuracy among the subtypes, see tab. 1, 2, 3. In fold 1 and 2 of the validation sets, no type Astrocytoma with the positive label was present, see tab. 7.

Similarly, for the type Oligodendroglioma, no negatively labelled case was present in fold 1, 2 and 3 in the validation set, see tab. 7.

3.1.1 Performance for 1000 Patches per Patient

The classification performance of the model using 1000 patches per patient was analysed for the training and test set separately. The mean Acc, AUCroc, and APS were reported including std, see tab. 4. There is a good agreement between the performance on the validation and test set which excludes the occurrence of overfitting. The accuracy on both sets reaches almost 70% while the APS is slightly higher for the validation set (81.0%, 7.3%) (mean APS, std) compared to the test set (76.3%, 5.8%), see tab. 4. The uncertainties are large (4.3%, 7.6%) (full range), see tab. 4.

	Acc	**AUCroc**	**APS**
validation set	68.6% ±4.3%	73.2% ±5.2%	81.0% ±7.3%
test set	67.2% ±7.3%	71.7% ±7.6%	76.3% ±5.8%

Table 4: Averaged Performance over 3 runs and 5 folds on validation and test set with 1000 patches per patient and the best hyper-parameters. The mean values including std are reported.

The loss curves of CV 1 and CV 2 show a different performance in terms of a bigger off-set between the validation and training curve in CV 2 compared to CV 1, see fig. 3, 4. CV 1 shows a good validation loss performance for the first 40 epochs while for the subsequent epochs the validation loss stops decreasing, see fig. 3. The area under the Roc curve for CV 1 (AUCroc = 0.81) is 12.3% larger than for CV 2 (AUCroc = 0.71), see fig. 6, 5. The area under the precision-recall curve for CV1 (AUPRC = 0.83) is 8.4% larger than for CV 2 (AUPRC = 0.76), see fig. 5, 6.

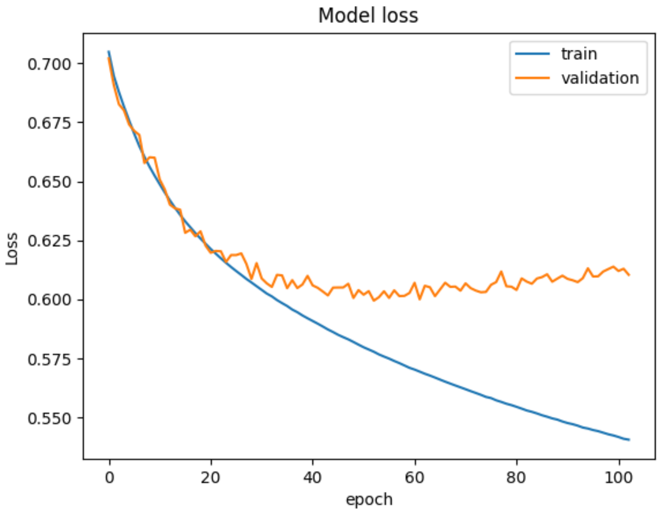


Figure 3: The loss curve of run 0/2 in CV 1 for 1000 patches shows a good validation loss performance for the first 40 epochs while for the subsequent epochs the problem of over-fitting arises as the distance between the training and validation loss increases over time.

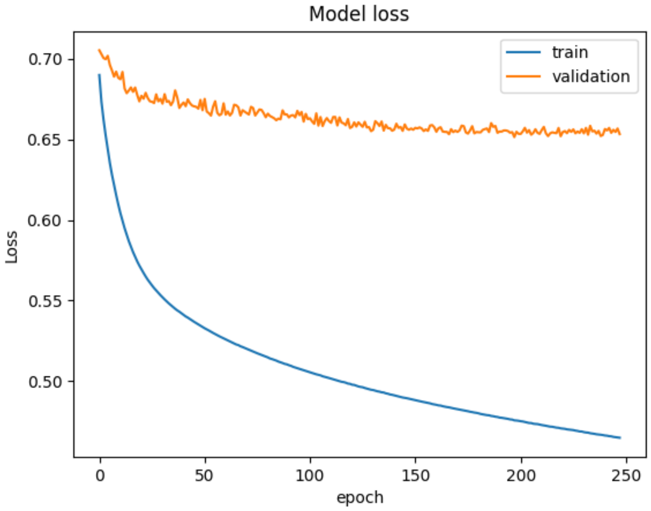


Figure 4: The loss curve of run 0/2 in CV 2 for 1000 patches shows a bigger off-set between the training and the validation curve than in CV 1, see fig. 3.

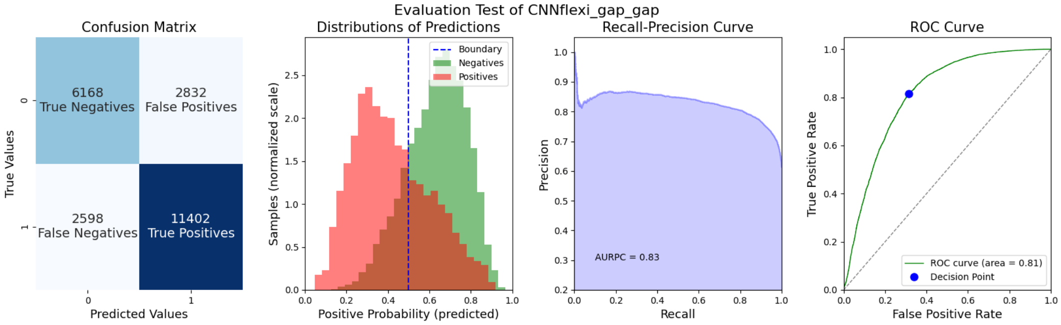


Figure 5: Overview of performance of run 0/2 of CV1 on the test set including the confusion matrix, the distributions of predictions with a threshold at 0.5, the recall-precision curve, and the ROC curve.

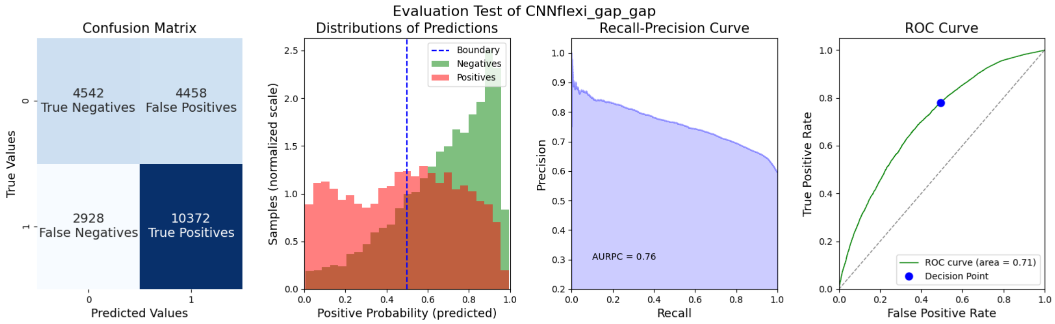


Figure 6: Overview of performance of run 0/2 of CV2 on the test set including the confusion matrix, the distributions of predictions with a threshold at 0.5, the recall-precision curve, and the ROC curve.

Effect of adapted validation set for CV2

Due to the fact that fold 2 of the cross-validation shows an aberrant curve for the validation loss, see fig. 4, further analysis was done for this subset. The validation set for CV 2 was checked for abnormalities in the patches. Quantitatively, the mean and the unit variance were checked as well as the intensity distribution in the patches were analysed. By visual inspection of the patches, no irregularities were observed compared to other patches of different patients. Also, no non-uniformity in the intensity distribution of the patches was observed. As a next step, the validation set of CV 2 was adapted by deleting two distinct patients at a time until every patient of the validation set was excluded once. The model was run for all adapted validation sets of CV2. After visually analysing the loss curves of the adapted validation sets, the best performing adapted validation set was selected. By excluding patients LGG-516 and LGG-391 the smallest off-set between the training and validation loss curve was observed, see fig. 7.

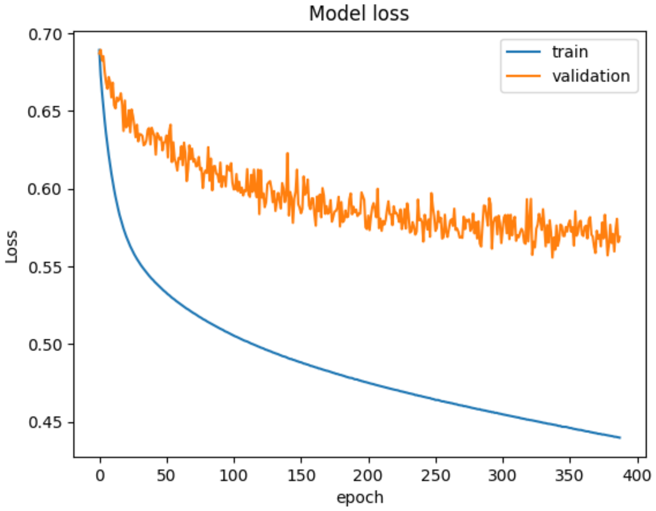


Figure 7: Loss curve for the best performing adapted validation set of CV2 excluding patient LGG-516 and LGG-391. The off-set between the validation and the training loss curve is smaller than when including all patients in fig. 4.

3.1.2 Performance for 60 Patches per Patient

The classification performance of the model using 60 patches per patient was analysed for the training and test set separately. The mean Acc, AUCroc, and APS were reported including std, see tab. 5. The uncertainties are larger when only including 60 patches per patient (4.8%, 12.4%) (full range) than when considering 1000 patches (4.3%, 7.6%), see tab. 4, 5.

	Acc	**AUCroc**	**APS**
validation set	67.0% ±4.8%	65.9% ±10.6%	76.4% ±10.2%
test set	64.8% ±8.0%	65.6% ±12.4%	72.5% ±10.0%

Table 5: Averaged Performance over 3 runs and 5 folds on validation and test set with 60 patches per patient and the best hyper-parameters. The mean values including std are reported.

A good agreement between the validation and training set can be observed, exluding the occurrence of overfitting, see tab. 5. The classification performance of the model on the validation set when considering 60 patches per patient (64.8 %, 8.0%) (mean accuracy, std) is slightly weaker than when considering 1000 patches per patient (67.2%, 7.3%) (mean accuracy, std). Given the large uncertainties in tab. 4, 5, the difference in performance when considering 60 vs. 1000 patches is not significant.

4 Discussion and Outlook

Discussion
In this preliminary study, a non-invasive method to predict 1p/19q chromosomal arm deletion from T2-FSE MR images using CNN approach was presented. As mentioned in previous studies, a subset of LGG are sensitive to therapy and have better survival chances, van den Bent et al. (2013), Macdonald et al. (1990), Franceschi et al. (2018). Given the limitations of histopathological diagnosis of tumors, adding other information such as 1p/19q deletion, which has been associated with positive response to therapy, could help improving therapeutic decision-making, Franceschi et al. (2018), Garcia et al. (2018), Ricard et al. (2007), van den Bent et al. (2013).

In order to reach a robust performance of the model, different hyperparamters can further be refined. Taking into account, that 94.6% patients contain more than 500 patches, 88.4% patients contain more than 1000 patches and 48.2% patients contain even more than 10000 patches, it would be interesting to see the performance of the model when the number of patches included per patient is augmented to e.g. 2000 patches. The differences between the patients in terms of number of patches included would then be more prominent, so that the patients would have to be weighted differently according to their available number of patches. Regarding the hyperparameter kernel size: wider kernels require more memory for parameter storage and increase runtime but a narrower output reduces memory cost. The consequence of a narrower output is the loss of details. As the image size is small (30x30) in this study as well there is an interest in micro structures of the tumor, one would not include a the kernel size larger than 4. The spikes visible in the validation loss curves, see fig. 3, 4, 7 are a sign for variation in performance. One could argue that the spikes are an unavoidable consequence of mini-batch Gradient Descent in the used Adam optimizer, ada. Some mini-batches have 'by chance' unlucky data for the optimization. A possible approach worth trying to fight the variation in the performance would be to try different optimizers. Akkus et al. (2017) evaluated the performance of different optimizers in the CNN used. SGD performed best in terms of Accuracy and Sensitivity (87.8%, 93.3%) compared to Adam (85.5%, 88.8%) and RMSprop (84.4%, 84.4%). Another idea is to replace the sigmoid activation function in the output layer by the softmax activation function or to replace the global average pooling layer by another dense layer. Once the performance of the model is stable and robust, additional convolutional layers can be added as with more layers, the architecture adapts to the high-level features.

The presented CNN reached 67.2% accuracy on the test set while Akkus et al. (2017) achieved an accuracy of 75.6% by only including T2 contrast images, without data augmentation. The reason for the lower performance in our case could be that we used 2D patches while Akkus et al. (2017) used 3D volumina. Further, Akkus et al. (2017) improved its accuracy to a value of 78.9% on the test set when combining T1 and T2 contrast images. The gain in performance is smaller than one would expect from the additional information of the soft tissue structure and potential patterns contained in different contrast images. In order to feed different image contrasts to the same model, the need of more layers and separate arms which increase the number of model parameters, arises. The comparison of our model to the literature demonstrates that the used CNN has to further be improved to reach higher performance values. Particularly, further investigations in the performance of CV 2 are necessary. The validation loss of fold 2, see fig. 4, shows a profoundly weaker performance than fold 1, see fig. 3, in terms of a big off-set. This was observed in all measurements including the case of 1000 patches per patient as well as when only using 60 patches per patient. Possible reasons for the better performing adapted validation set, see fig. 7, compared to the full validation set in CV2, see fig. 4, can be underlying unknown confounders which are more present in certain patients. The study of Zhao et al. (2020) emphasizes the presence of confounding effects as one of the most critical challenges in using deep learning in medical imaging studies. Confounders affect the relationship between input data (e.g., brain MRIs) and output variables (e.g., diagnosis), Zhao et al. (2020). Possible confounders could be potential brain edema present, individual tumor progression, internal surgery scars, sex, and age. Possible methods to overcome the problem of confounders are stratification and restriction of data, Zhao et al. (2020). The method of restricting the data was applied to overcome the different performance between CV 1 and CV 2 by excluding some patients. Stratification is defined as the act of sorting data into distinct groups. Patients could be stratified by label and tumor type.

A possible reason why the different folds are performing dissimilar to a large extend might be the unbalanced distribution of the folds due to the limited and unbalanced dataset, see tab. 1, 2, 6, 7, 8, as well as the mentioned unknown confounding factors present.

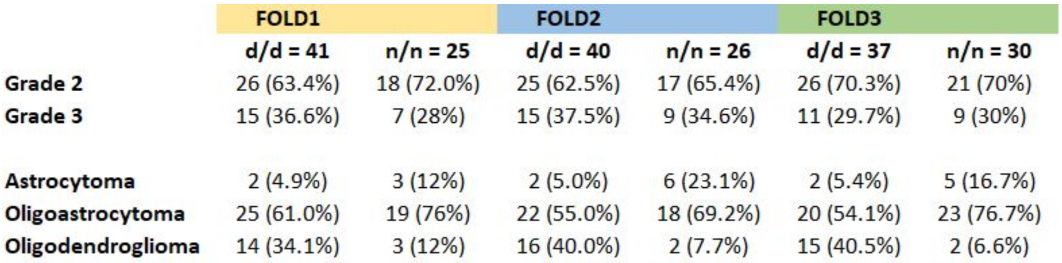
To overcome the problem of overfitting, the dataset should be augmented using the ImageDataGenerator from TensorFlow by randomly rotating, shifting and flipping the data. Akkus et al. (2017) reached an accuracy of 87.7% when the best configuration, e.g., the combination of T1 and T2 images including data augmentation and further training, was applied.

Outlook

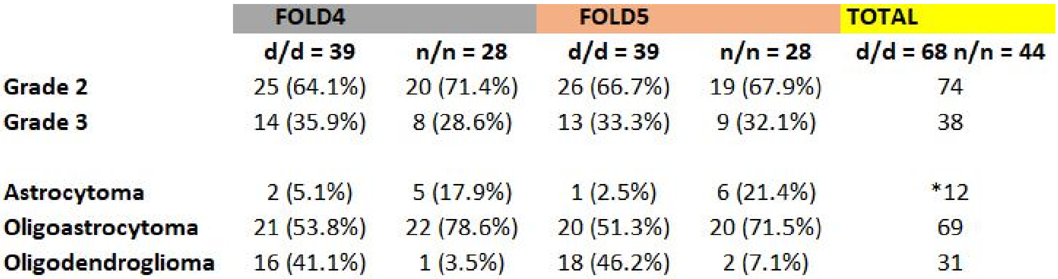
The aim of this lab rotation was the prediction of 1p/19q status from FSE-T2 MR images using a convolutional neural network (CNN). The performance of the presented model is definitely not good enough yet for clinical use. Particularly, the accuracy for the underrepresented tumor type Astrocytoma reached an accuracy of $\sim 53\%$ on the test set. To further examine the differences in the performance of the folds, one could redo the partitions and find out what patients cause the weaker performance in the folds and investigate in those differences between patients and thereby uncover the underlying confounders. Additionally, visual explanations in form of an interpretability analysis could help to improve the model. The approach, called Gradient-weighted Class Activation Mapping (Grad-CAM), is a technique that makes CNN-based models more transparent by visualizing the regions of input that are important for prediction, Selvaraju et al. (2017). The knowledge of which tumor patches carry the most classification benefit and where in the brain they are localised would help to improve the model.

Further studies with larger patient populations are required to investigate these and confirm the current findings. Once the presented method is improved it could potentially be used as an alternative to surgical biopsy and pathological analysis for predicting 1p/19q co-deletion status.

5 Appendix

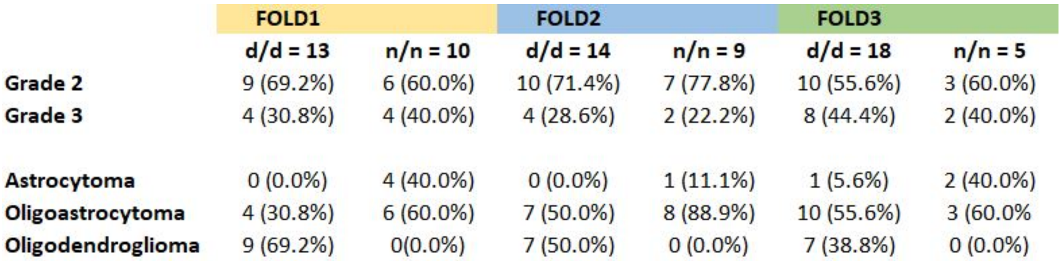


	FOLD1		FOLD2		FOLD3	
	d/d = 41	n/n = 25	d/d = 40	n/n = 26	d/d = 37	n/n = 30
Grade 2	26 (63.4%)	18 (72.0%)	25 (62.5%)	17 (65.4%)	26 (70.3%)	21 (70%)
Grade 3	15 (36.6%)	7 (28%)	15 (37.5%)	9 (34.6%)	11 (29.7%)	9 (30%)
Astrocytoma	2 (4.9%)	3 (12%)	2 (5.0%)	6 (23.1%)	2 (5.4%)	5 (16.7%)
Oligoastrocytoma	25 (61.0%)	19 (76%)	22 (55.0%)	18 (69.2%)	20 (54.1%)	23 (76.7%)
Oligodendroglioma	14 (34.1%)	3 (12%)	16 (40.0%)	2 (7.7%)	15 (40.5%)	2 (6.6%)

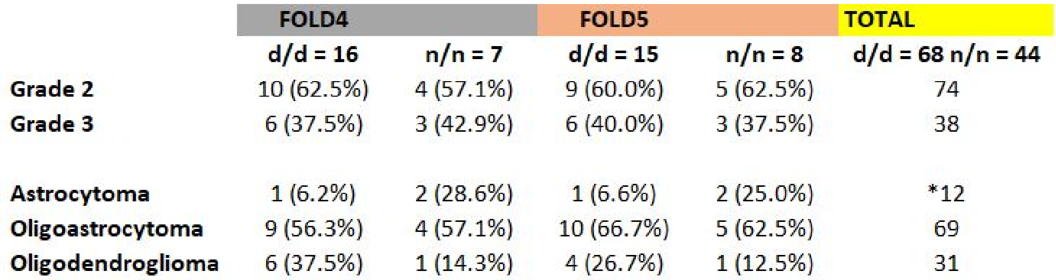


	FOLD4		FOLD5		TOTAL
	d/d = 39	n/n = 28	d/d = 39	n/n = 28	d/d = 68 n/n = 44
Grade 2	25 (64.1%)	20 (71.4%)	26 (66.7%)	19 (67.9%)	74
Grade 3	14 (35.9%)	8 (28.6%)	13 (33.3%)	9 (32.1%)	38
Astrocytoma	2 (5.1%)	5 (17.9%)	1 (2.5%)	6 (21.4%)	*12
Oligoastrocytoma	21 (53.8%)	22 (78.6%)	20 (51.3%)	20 (71.5%)	69
Oligodendroglioma	16 (41.1%)	1 (3.5%)	18 (46.2%)	2 (7.1%)	31

Table 6: Distribution of tumor label (d/d = double-deleted chromosomal arms, n/n = non-deleted chromosomal arms of 1p/19q), type, and grade in the training sets. The percentages were calculated based on the total number of patients within the label for grades and types separately. The subtype Astrocytoma (*) was represented in clear minority.

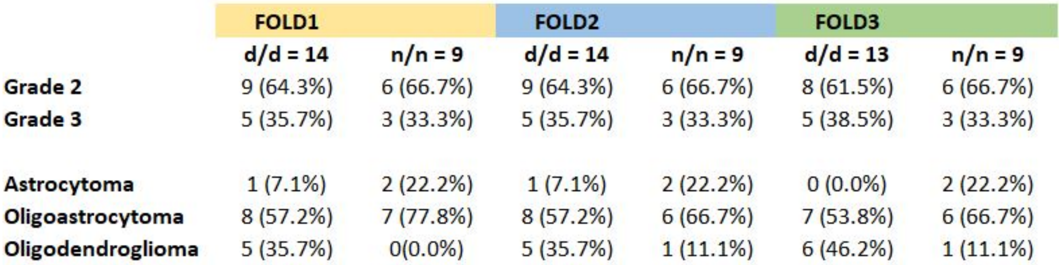


	FOLD1		FOLD2		FOLD3	
	d/d = 13	n/n = 10	d/d = 14	n/n = 9	d/d = 18	n/n = 5
Grade 2	9 (69.2%)	6 (60.0%)	10 (71.4%)	7 (77.8%)	10 (55.6%)	3 (60.0%)
Grade 3	4 (30.8%)	4 (40.0%)	4 (28.6%)	2 (22.2%)	8 (44.4%)	2 (40.0%)
Astrocytoma	0 (0.0%)	4 (40.0%)	0 (0.0%)	1 (11.1%)	1 (5.6%)	2 (40.0%)
Oligoastrocytoma	4 (30.8%)	6 (60.0%)	7 (50.0%)	8 (88.9%)	10 (55.6%)	3 (60.0%
Oligodendroglioma	9 (69.2%)	0(0.0%)	7 (50.0%)	0 (0.0%)	7 (38.8%)	0 (0.0%)

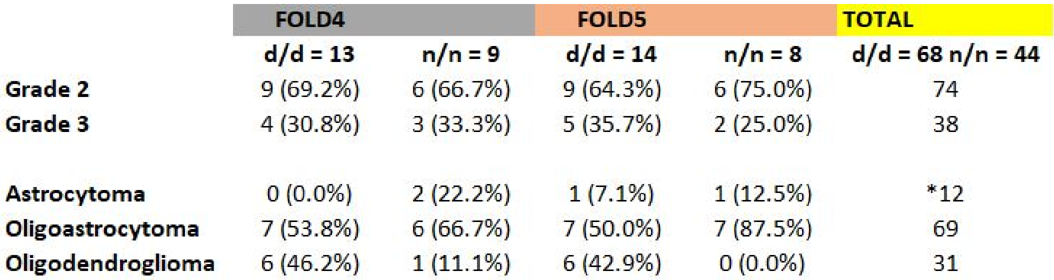


	FOLD4		FOLD5		TOTAL
	d/d = 16	n/n = 7	d/d = 15	n/n = 8	d/d = 68 n/n = 44
Grade 2	10 (62.5%)	4 (57.1%)	9 (60.0%)	5 (62.5%)	74
Grade 3	6 (37.5%)	3 (42.9%)	6 (40.0%)	3 (37.5%)	38
Astrocytoma	1 (6.2%)	2 (28.6%)	1 (6.6%)	2 (25.0%)	*12
Oligoastrocytoma	9 (56.3%)	4 (57.1%)	10 (66.7%)	5 (62.5%)	69
Oligodendroglioma	6 (37.5%)	1 (14.3%)	4 (26.7%)	1 (12.5%)	31

Table 7: Distribution of tumor label (d/d = double-deleted chromosomal arms, n/n = non-deleted chromosomal arms of 1p/19q), type, and grade in the validation sets. The percentages were calculated based on the total number of patients within the label for grades and types separately. The subtype Astrocytoma (*) was represented in clear minority.



	FOLD1		FOLD2		FOLD3	
	d/d = 14	n/n = 9	d/d = 14	n/n = 9	d/d = 13	n/n = 9
Grade 2	9 (64.3%)	6 (66.7%)	9 (64.3%)	6 (66.7%)	8 (61.5%)	6 (66.7%)
Grade 3	5 (35.7%)	3 (33.3%)	5 (35.7%)	3 (33.3%)	5 (38.5%)	3 (33.3%)
Astrocytoma	1 (7.1%)	2 (22.2%)	1 (7.1%)	2 (22.2%)	0 (0.0%)	2 (22.2%)
Oligoastrocytoma	8 (57.2%)	7 (77.8%)	8 (57.2%)	6 (66.7%)	7 (53.8%)	6 (66.7%)
Oligodendroglioma	5 (35.7%)	0(0.0%)	5 (35.7%)	1 (11.1%)	6 (46.2%)	1 (11.1%)



	FOLD4		FOLD5		TOTAL
	d/d = 13	n/n = 9	d/d = 14	n/n = 8	d/d = 68 n/n = 44
Grade 2	9 (69.2%)	6 (66.7%)	9 (64.3%)	6 (75.0%)	74
Grade 3	4 (30.8%)	3 (33.3%)	5 (35.7%)	2 (25.0%)	38
Astrocytoma	0 (0.0%)	2 (22.2%)	1 (7.1%)	1 (12.5%)	*12
Oligoastrocytoma	7 (53.8%)	6 (66.7%)	7 (50.0%)	7 (87.5%)	69
Oligodendroglioma	6 (46.2%)	1 (11.1%)	6 (42.9%)	0 (0.0%)	31

Table 8: Distribution of tumor label (d/d = double-deleted chromosomal arms, n/n = non-deleted chromosomal arms of 1p/19q), type, and grade in the test sets. The percentages were calculated based on the total number of patients within the label for grades and types separately. The subtype Astrocytoma (*) was represented in clear minority.

6 Bibliography

A. P. Patel, J. L. Fisher, E. Nichols, F. Abd-Allah, J. Abdela, A. Abdelalim, H. N. Abraha, D. Agius, F. Alahdab, T. Alam *et al.*, "Global, regional, and national burden of brain and other cns cancer, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 4, pp. 376–393, 2019.

S. Cha, "Update on brain tumor imaging: From anatomy to physiology," *American Journal of Neuroradiology*, vol. 27, no. 3, pp. 475–487, 2006.

P. Kleihues, P. C. Burger, and B. W. Scheithauer, "The new who classification of brain tumours," *Brain pathology*, vol. 3, no. 3, pp. 255–268, 1993.

C. G. A. R. Network, "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015.

M. J. van den Bent, A. A. Brandes, M. J. Taphoorn, J. M. Kros, M. C. Kouwenhoven, J.-Y. Delattre, H. J. Bernsen, M. Frenay, C. C. Tijssen, W. Grisold *et al.*, "Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of eortc brain tumor group study 26951," *Journal of clinical oncology*, vol. 31, no. 3, pp. 344–350, 2013.

D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, "The 2007 who classification of tumours of the central nervous system," *Acta neuropathologica*, vol. 114, no. 2, pp. 97–109, 2007.

D. R. Macdonald, L. E. Gaspar, and J. G. Cairncross, "Successful chemotherapy for newly diagnosed aggressive oligodendroglioma," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 27, no. 5, pp. 573–574, 1990.

E. Franceschi, A. Mura, D. De Biase, G. Tallini, A. Pession, M. P. Foschini, D. Danieli, S. Pizzolitto, E. Zunarelli, G. Lanza *et al.*, "The role of clinical and molecular factors in low-grade gliomas: what is their impact on survival?" *Future Oncology*, vol. 14, no. 16, pp. 1559–1567, 2018.

C. R. Garcia, S. A. Slone, T. Pittman, W. H. St. Clair, D. D. Lightner, and J. L. Villano, "Comprehensive evaluation of treatment and outcomes of low-grade diffuse gliomas," *PloS one*, vol. 13, no. 9, p. e0203639, 2018.

D. Ricard, G. Kaloshi, A. Amiel-Benouaich, J. Lejeune, Y. Marie, E. Mandonnet, M. Kujas, K. Mokhtari, S. Taillibert, F. Laigle-Donadey *et al.*, "Dynamic history of low-grade gliomas before and after temozolomide treatment," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 61, no. 5, pp. 484–490, 2007.

D. Scheie, P. A. Andresen, M. Cvancarova, A. S. Bø, E. Helseth, K. Skullerud, and K. Beiske, "Fluorescence in situ hybridization (fish) on touch preparations: a reliable method for detecting loss of heterozygosity at 1p and 19q in oligodendroglial tumors," *The American journal of surgical pathology*, vol. 30, no. 7, pp. 828–837, 2006.

C.-d. L.-g. Quang-Hien Kha, "Development and Validation of an Efficient MRI Radiomics Signature for Improving the Predictive Performance of 1p / 19q," pp. 1–15, 2021.

S. Fellah, D. Caudal, A. M. De Paula, P. Dory-Lautrec, D. Figarella-Branger, O. Chinot, P. Metellus, P. J. Cozzone, S. Confort-Gouny, B. Ghattas *et al.*, "Multimodal mr imaging (diffusion, perfusion, and spectroscopy): is it possible to distinguish oligodendroglial tumor grade and 1p/19q codeletion in the pretherapeutic diagnosis?" *American Journal of Neuroradiology*, vol. 34, no. 7, pp. 1326–1333, 2013.

N. L. Jansen, C. Schwartz, V. Graute, S. Eigenbrod, J. Lutz, R. Egensperger, G. Pöpperl, H. A. Kretzschmar, P. Cumming, P. Bartenstein *et al.*, "Prediction of oligodendroglial histology and loh 1p/19q using dynamic [18f] fet-pet imaging in intracranial who grade ii and iii gliomas," *Neuro-oncology*, vol. 14, no. 12, pp. 1473–1480, 2012.

Y. Iwadate, N. Shinozaki, T. Matsutani, Y. Uchino, and N. Saeki, “Molecular imaging of 1p/19q deletion in oligodendroglial tumours with 11c-methionine positron emission tomography,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 87, no. 9, pp. 1016–1021, 2016.

L. M. DeAngelis, “Brain tumors,” *New England journal of medicine*, vol. 344, no. 2, pp. 114–123, 2001.

Z. Akkus, I. Ali, J. Sedlář, J. P. Agrawal, I. F. Parney, C. Giannini, and B. J. Erickson, “Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence,” *Journal of Digital Imaging*, vol. 30, no. 4, pp. 469–476, 2017.

S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

“LGG-1p19qDeletion - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki,” https://wiki.cancerimagingarchive.net/display/Public/LGG-1p19qDeletion, accessed: 2021-10-11.

A. J. Sankowski, J. B. Ćwikła, M. L. Nowicki, S. Chaberek, M. Pech, A. Lewczuk, and J. Walecki, “The clinical value of mri using single-shot echoplanar dwi to identify liver involvement in patients with advanced gastroenteropancreatic-neuroendocrine tumors (gep-nets), compared to fse t2 and ffe t1 weighted image after iv gd-eob-dtpa contrast enhancement,” *Medical science monitor: international medical journal of experimental and clinical research*, vol. 18, no. 5, p. MT33, 2012.

“sklearn,” https://scikit-learn.org/stable/, accessed: 2021-10-16.

“PyPl - patchify,” https://pypi.org/project/patchify.html, accessed: 2021-10-20.

H. H. Aghdam and E. J. Heravi, “Guide to convolutional neural networks,” *New York, NY: Springer*, vol. 10, no. 978-973, p. 51, 2017.

S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, “Residual and plain convolutional neural networks for 3d brain mri classification,” in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE, 2017, pp. 835–838.

S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*. Ieee, 2017, pp. 1–6.

T. Guo, J. Dong, H. Li, and Y. Gao, “Simple convolutional neural network on image classification,” in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 721–724.

X. Lin, C. Zhao, and W. Pan, “Towards accurate binary convolutional neural network,” *arXiv preprint arXiv:1711.11294*, 2017.

“Keras: the Python deep learning API,” https://keras.io/, accessed: 2021-21-10.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

Q. Li, M. Yan, and J. Xu, “Optimizing convolutional neural network performance by mitigating underfitting and overfitting,” in *2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)*. IEEE, 2021, pp. 126–131.

“A detailed example of data generators with Keras,” https://stanford.edu/{~}shervine/blog/keras-how-to-generate-data-on-the-fly, accessed: 2021-11-01.

I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

"tf.keras.optimizers.Adam | TensorFlow Core v2.7.0," https://www.tensorflow.org/api{_}docs/python/tf/keras/optimizers/Adam, accessed: 2021-11-16.

Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.