Machine Learning for Healthcare: Project 3

Mert Ertugrul, Johan Lokna, Nora Schneider

I. Introduction

Brain cancer is one of the most life-threatening cancer types. A common approach for detecting brain tumors is by MRI (Magnetic Resonance Imaging) scans. Usually, these images are examined by radiologists. However, a manual inspection can be time-consuming and error-prone, which motivates the need for automated tools [5]. Meanwhile, there is an increasing interest in explainable AI models, specifically in the medical domain. According to the GDPR, every individual has the right to a meaningful explanation for automated decision-making. Further, ethical considerations, performance increase and risk reduction are some drivers of this increasing trend [3]. In this report we evaluate different machine learning approaches for detecting brain tumors. Furthermore, we also apply different post-hoc explanation methods on each model to evaluate their interpretability.

A. Dataset

We implement and validate our models based on the provided dataset [2, 4]. It consists of MRI brain scans with and without tumors. Due to duplicates and the resulting risk of information leakage, we first filter the data. We consider duplicates arising from rotation, shifting and scaling. Still, we provide no guarantees that all duplicate pairs have been found and removed. The filtered dataset contains 87 and 71 images with and without a tumor. As the relative small size of the dataset poses a challenge for most standard machine learning models, we apply transfer learning as a possible remedy, for which we use an additional dataset [6]. It includes $1,500$ tumorous and tumor-free images, respectively.

In order to improve the quality of our models, we apply techniques from image analysis to enhance the brain scans. Firstly, we remove the scull. Then, we ensure that the intensity distribution is similar across images by means of histogram equalization [16]. We also tested scaling images so that the brain scans to fill the entire picture and also subtracting the most dominant PCA-components; this did not effect the quality substantially ans is therefore not used. Further to increase the dataset size we used standard augmentation methods such as random rotation.

Additionally, a preprocessed dataset with extracted features by pyradiomics was given. With our given resources, it is not possible to filter duplicates for the pyradiomics dataset. Hence, corresponding results need to be considered carefully.

II. Feature-based Approach

We train a random forest classifiers on the pyradiomics features and carry out hyperparameter grid search. We also test feature selection, but the model performs better with the full set of features. The results of this section cannot be compared fairly with the CNN based results, as the image dataset is duplicate-free. After obtaining a final random forest baseline model, we apply the following post-hoc interpretability methods and observe the most influential features:

MDI and MDA: MDI uses the mean decrease in GINI impurity. Assuming independent input variables it can correspond to a variance decomposition of the output [14]. However, this assumption is far fetched for the given high dimensional feature space. MDA or Mean Decrease Accuracy uses random permutations of the feature values to evaluate feature importance. This method assigns noticeable importance to only 4 features, which are highly general features of wavelet and first order type. In contrast MDI distributes importance among a larger subset of the features, although it again ranks similar features as the most important ones.

SHAP vs MDI: We obtain the global SHAP values on our test set. The values correspond to the marginal contribution of each pyrodiomics feature averaged over permutations of features. Compared to the MDI feature importances, SHAP importances rank the top 10 features in identical order. The most important features are consistently in the following groups: first order features (10 percentile, median, wavelet etc.) and shape related features such as perimeter-surface ratio. The importance of first order features is intuitive as they express statistics of voxel/pixel intensity distribution and tumors tend to have a lighter pixel value in our training set. Overall, the interpretation suggests that this model uses color distribution and shape characteristics to deduce that there is a visual change corresponding to a tumor.

Lime: Local interpretable model agnostic explanations (LIME) is used to explain predictions of black box machine learning models on individual data points. LIME works by training "surrogate models" that aim to approximate the original ml model's prediction on an individual data point, using a new dataset created by perturbations

to the intended data point and how they affect model classification [11, 12].

Compared to MDI or the SHAP based feature importances, LIME's importance assignment to features is less informative. For our dataset and a random forest classifier, whose structure can readily be leveraged to construct an interpretation, LIME is not on par with the other methods used.

III. CNN-based Approach

We implement and compare different CNN based models as they are well-suited for image data. We compare their performances using accuracy and f1-scores. Subsequently, we interpret all models with SHAP, GradCam and integrated gradients.

A. Models

Standard CNN: We implement the CNN baseline, which has a total of $1,606,802$ parameters. We tune hyperparameters using a grid search and finally train the classifier on the original data and on an augmented version of it. In order to reduce complexity and avoid overfitting, we further adjust the architecture by introducing dropout and maxpooling layers, resulting in a CNN with $294,818$ parameters. Last, as a possible solution for the small data set regime, we use the same adjusted architecture with transfer learning. We first train the CNN on the large data [6] and then retrain it on the original data.

Variational Autoencoder: We work with a variational autoencoder (VAE) as a potentially more interpretable construction method for a final CNN based classifier. The idea is that a VAE is trained on our image dataset to learn a latent space representation to reconstruct given images. Our expectation is that this process would produce convolution filters in the encoder that are able to capture structural information from the image [8]. By working with a VAE instead of a standard autoencoder, our aim is to enforce more localised representations of structural features in the latent space [1, 8]. After VAE training, we extract the encoder section and combine it with fully connected layers to create a classifier, training again with the pretrained weights of the encoder. Our expectation is that initialising the classifier with a localised latent space distribution would improve interpretability.

When only the dataset of the project is used, the VAE cannot converge in a meaningful way as high frequency and complex features of brain images are not successfully reconstructed from a small dataset. Therefore, we opt for transfer learning by training the VAE and later the final classifier on a larger but similar dataset [6]. Then, we fine tune the final classifier with the project dataset.

UNet: For many medical applications, it is not sufficient to only determine the presence of a brain tumor; often, one must also locate the cancerous region. Moreover, if users can visually observe that the model correctly identifies tumorous sections, it demystifies the prediction process and increases confidence in the model. In order to pursue this more ambitious goal, we train a segmentation model based on U-Net, which is known to converge even for small sample sizes [13].

We want the segmentation mask to indicate whether tissue is healthy or cancerous, which is generally a supervised learning task. Therefore, we manually annotate the images indicating which regions are tumorous according to our non-expert eyes. Class imbalance caused by the dominance of healthy tissue in the brain scans is tackled by using a weighted average between the cross-entropy and dice loss [7].

As for the previous models, we want to utilize the larger dataset [6] for increasing the model's performance. Although the dataset is unlabeled, we can use semi-supervised techniques to improved the model's consistency [9, 18]. In order to obtain a gradient signal from an unlabeled image, we inject noise and use the cross-entropy between the perturbed and unperturbed predictions as loss function. This allows us to use unlabeled images to regularize the model and thereby improve its consistency.

After training the segmentation model, we construct a classifier which predicts that a brain scan contains a tumor if the segmentation model indicates that any region is cancerous with a probability over a certain threshold, which is learned on the validation set. This allows the model to remain intuitive and interpretive.

B. Performances

Standard CNN: The baseline CNN is not very robust, as indicated by its high standard deviation upon repeated training. Performance and variance can be slightly improved by using augmented data. However, the model seems to be too complex for the few data points provided. We therefore suspect variance to be the dominating factor in the bias-variance trade-off. Applying the less complex, advanced CNN improves the accuracy and reduces the variance. Yet, performance is still not stable.

Using the advanced CNN architecture with transfer learning performs best among all methods. This shows the capability of the adjusted CNN architecture to learn in a meaningful way, if it has enough data to train on. It is worth mentioning, that model performance is the same for retraining the entire network or only fully connected layers. Inspecting the single wrongly classified image in the test set, one observes that the tumorous region is of much lower intensity than in other images. This can be considered an out of distribution sample compared

to the training data, which is possibly why the classifier consistently unable to correctly classify it.

Variational Autoencoder: Although the variational autoencoder uses the larger transfer learning dataset, the corresponding classifier performs similar to the advanced CNN without transfer learning. Some convergence is observed during transfer learning, while this is not the case during fine tuning. The structure and high frequency patterns of the brain may be too complex for our autoencoder to learn, hence the pretrained encoder cannot give the classifier a head start for fine tuning. A VAE with higher capacity or a conditional GAN would be a more suitable alternative for future work. Image reconstruction with local perturbations in the embedding space shows only simple patterns as a result see section V-E.

UNet: A common metric for evaluating a segmentation model is its dice score. The UNet model trained exclusively on labeled images, attains a score of 0.266 ± 0.032. Furthermore, when including unlabeled images for consistency regularization, the score improves to 0.324 ± 0.005. This indicates that the unregularized model's predictions might be overly spurious, which is symptomatic to the small sample size regime.

The unregularized UNet-based model attains the highest accuracy and f1-score for any model trained exclusively on the original dataset, as seen in Table I. Moreover, the improvement in the underlying segmentation model through consistency regularization is reflected by the quality of the classifier; accuracy and f1-score improve by 1.5% and 2% respectively. However, the UNet-model benefits significantly less than the advanced CNN from the large dataset. This highlights a limitation to this approach; data acquisition is significantly more involved. However, when such data is available, this approach works well.

Model	Accuracy	F1-Score
Random Forest	0.7857	0.8500
Baseline CNN	0.7049 ± 0.0973	0.7070 ± 0.1268
Baseline CNN + Augmentation	0.7250 ± 0.0750	0.7150 ± 0.0952
Advanced CNN	0.7400 ± 0.0663	0.7592 ± 0.0824
Advanced CNN + TL	**0.9500 ± 0.0000**	**0.9565 ± 0.0000**
VAE + TL	0.7500	0.7619
UNet	0.8500 ± 0.0591	0.8564 ± 0.0718
UNet + Unlabeled	0.8650 ± 0.0440	0.8779 ± 0.0376

TABLE I: Model Performances; the largest value for each column is highlighted, for highly variable models mean and standard-deviation is reported.

C. Interpretability

We evaluate different approaches to explain our models' decision making: SHAP [10], GradCam [15] and IntegratedGradients [17]. A representative visual summary of results can be seen in Figure 1 and 2. Here, the model considers blue colored regions as evidence for a tumor while red colored regions indicates the opposite. In order to compare the explainability methods, we consider the perceived region of influence according to the different techniques. Table II displays the average Intersect-over-Union (IoU) between this region and the annotated segmentation mask used for training the UNet model. For any tumorous observation, we suspect that a well calibrated model will put emphasis on the cancerous region in order to determine the presence of a tumor. Therefore, we conjecture that the IoU-metric will be a good measure for a model's explainability given a certain interpretabiliity method.

SHAP: SHAP assigns each pixel an importance value for the corresponding prediction output, while not evaluating the quality of the prediction itself. Therefore looking at the SHAP values for an input image can explain which regions strongly contribute to the prediction.

For the baseline CNNs, the SHAP values of the prediction are very spurious, as can be seen in Figure 1. Generally, the model puts emphasis on the tumorous region as well as the scull, when it's not removed. Moreover, the boundary between light and dark regions seems to be of interest for the classifier. However, there does not seem to exist any regions which are solely associated with a specific prediction. This is then reflected by the relative low IoU scores. In total, SHAP values are consistent with the baseline models' performance, as it also has a high variance and a low accuracy.

Other tested models are more interpretable according to their SHAP-values; we observe a stronger overlap between the model's region of influence and the tumorous section. In figure 1 we see that all of them identify the tumor as a region of interest. Yet, the advanced CNN and UNet + unlabeled obtain significantly highest IoU scores. For the tumor-free sample in Figure 2 there are no larger regions indicating a positive correlation with the output "tumor". Instead the brain structure itself seems to have a negative correlation, which can be nicely seen for the advanced CNN in figure 1.

Integrated Gradients: is another gradient based post-hoc interpretability method. Given an input image, this method obtains the gradient of the input image itself with respect to the final loss instead of an intermediary convolutional layer as is done for GradCam. Therefore, this method produces an interpretation of the whole model instead of a section of it. [17]

The IoU scores in table II show, that the UNet's gradients provide a highly localised importance region that matches our manually labelled tumor segment very closely. After UNet, the advanced CNN is next in line

according to our metric. Inspecting the sample visuals in figure 1, we see that indeed most models intuitively highlight the correct tumor region. Still, the VAE generally seems focus more on the center of the image compared to the tumours region, which might explain its convergence issues.

Grad Cam: Given an input image and a convolutional layer GradCam analyzes the output of the layer and find its gradient with respect to the final model loss. Then, this gradient is rescaled to be overlaid on the input image for visual inspection [15].

GradCam interpretability results highly depend on the chosen layer and model architecture. Therefore, comparing different models is difficult. Analyzing the IoU scores, the GradCam mask from variational autoencoder is best. Generally, GradCam has higher IoU scores then SHAP. Visually, GradCam is the only method that appears intuitive for the VAE based model on the sample image in 1 while integrated gradients appears more intuitive for UNet. We deduce that it is good practice to try a number of methods at once when interpreting a result, as their success is dependent on model structure too.

Model	**SHAP**	**Integrated Gradient**	**GradCam**
Baseline CNN	0.0711	0.0815	0.0370
Baseline CNN + Augmentation	0.0913	0.0873	0.2043
Advanced CNN	**0.3190**	0.3107	0.2718
Advanced CNN + Transfer Learning	0.2239	0.1553	0.3596
VAE based CNN + Transfer Learning	0.0942	0.0958	**0.4394**
UNet	0.2153	**0.5684**	0.0760
UNet + Unlabeled	0.3142	0.5637	0.3775

TABLE II: Intersect-over-Union for tumor samples in the test set; We compare the mask obtained by the interpretability model with the true segmentation used for the UNet model

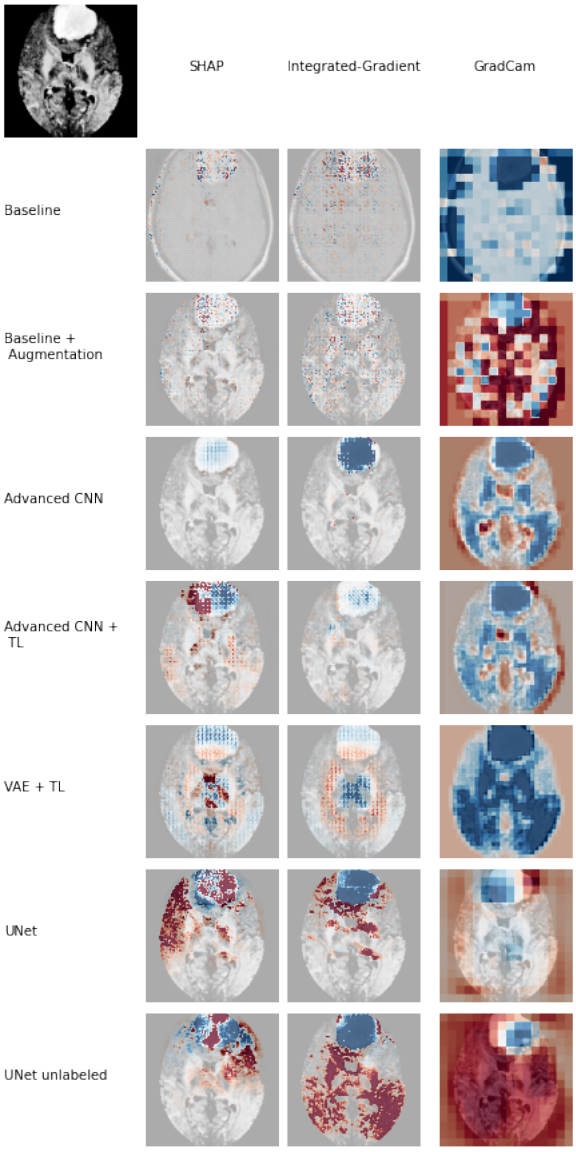

Fig. 1: Interpretability on a test sample with a tumor; Blue pixels positively correlate with predicting ”tumor”, while red pixels indicate negative correlation

IV. Conclusion

We test and evaluate different state-of-the-art machine learning models for detecting tumors in MRI brain scans. Further, we interpret the models by applying state of the art interpretability methods. The relative small dataset size is a challenge that resulted in all of our models being spurious. The variance of model performance can be reduced by using an additional larger dataset.

Comparing the different interpretability methods (SHAP, Integrated Gradient, GradCam) by IoU scores, there is no approach that performs best for all models. However, there is a trend that integrated gradient and GradCam have higher IoU scores than SHAP.

Performance-wise the advanced CNN with transfer learning outperforms any other model. However, using the IoU as a measure of interpretability, the model is worse than its non-transfer learning counter part with respect to all examined interpretability methods and the UNet. This is consistent with a visual inspection of interpretability plots. However, the second best performing model, UNet paired with using the unlabeled data from the transfer-learning dataset, obtains best IoU scores.

When no additional data is used, UNet performs best among all models and is still very interpretable as indicated by IoU values. Therefore, in a performance-interpretability tradeoff we conclude that UNet with optionally using additional data in a semi-supervised way is the best model for the classification task.

REFERENCES

[1] Chitresh Bhushan, Zhaoyuan Yang, Nurali Virani, and Naresh Iyer. Variational encoder-based reliable classification. *CoRR*, abs/2002.08289, 2020.

[2] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri), 2020.

[3] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *CoRR*, abs/2102.13076, 2021.

[4] Navoneel Chakrabarty. Brain mri images for brain tumor detection, 2019.

[5] S. Deepak and P.M. Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in Biology and Medicine*, 111:103345, 2019.

[6] Ahmed Hamad. Br35h :: Brain tumor detection 2020, 2020.

[7] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, oct 2020.

[8] Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167, 2015.

[9] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation, 2021.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[11] Christoph Molnar. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/, 2018. https://christophm.github.io/interpretable-ml-book/.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[14] Erwan Scornet. Trees, forests, and impurity-based variable importance, 2020.

[15] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[16] N. Senthilkumaran and J. Thimmiaraja. Histogram equalization for image enhancement using mri brain images. In *2014 World Congress on Computing and Communication Technologies*, pages 80–83, 2014.

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[18] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2019.

V. Appendix

A. Interpretability of models on a normal sample

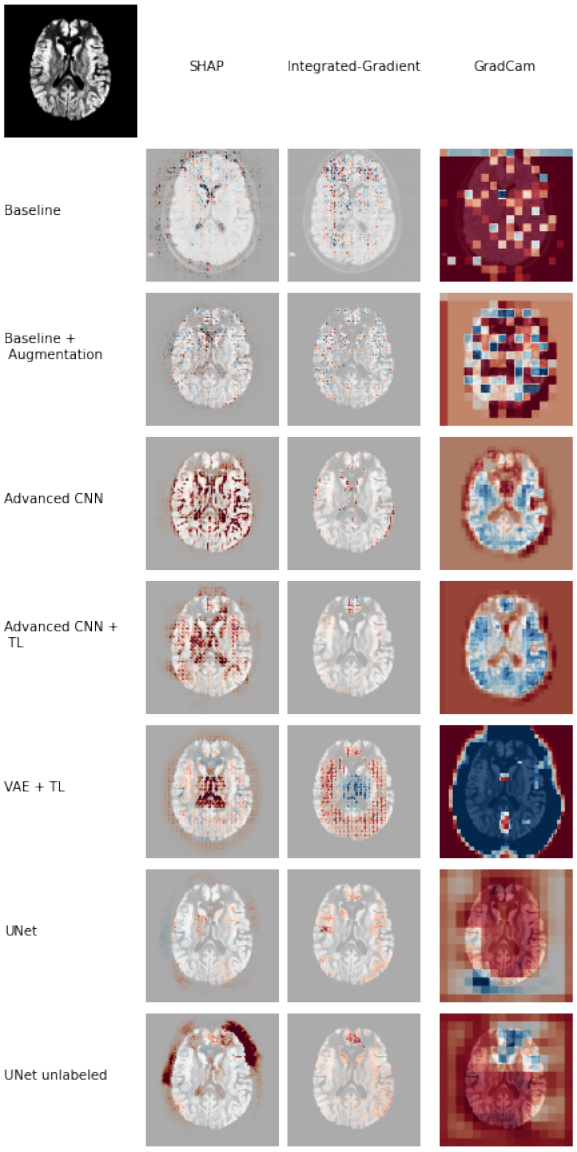

Fig. 2: Interpretability on a test sample without a tumor; Blue pixels positively correlate with predicting "tumor", while red pixels indicate negative correlation

B. Random Forest MDI Feature Weights

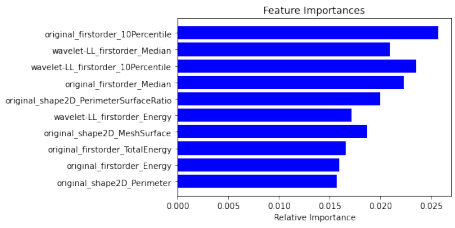

Fig. 3: Gini Impurity Based Feature Importances

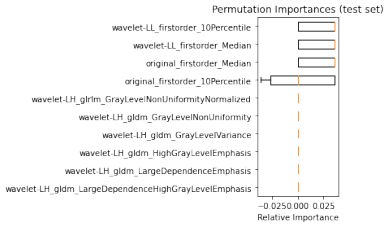

Fig. 4: Permutation Based Feature Importances

C. Random Forest MDA Feature Weights

D. Random Forest Shaply Features

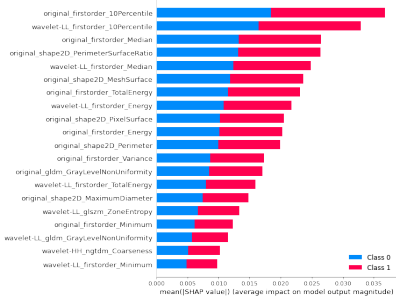

Fig. 5: Shaply Based Feature Importances

E. Variational Autoencoder Perturbed Reconstructions

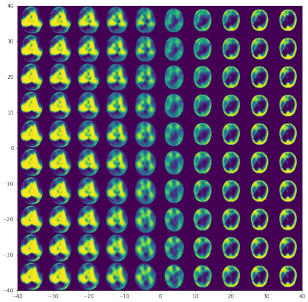

Fig. 6: Reconstruction from latent space Perturbations on a sample image

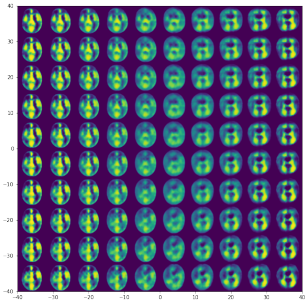

Fig. 7: Reconstruction from latent space Perturbations on a sample image