Machine Learning for Healthcare: Interpretable Medical Image Classification

Assignment 3

Arka Mitra
amitra@student.ethz.ch

Maximilian Hildebrandt
mhildebrandt@student.ethz.ch

May 17, 2022

1 Introduction

Growing amounts of medical equipment are available (Swiss Federal Bureau of Statistics, 2021), which produces a wealth of data also suitable for machine learning applications. This report aims to apply a range of interpretable and explainable machine learning techniques to the task of tumor identification on the Kaggle Brain Tumor dataset (Kaggle, n.d.). Specifically, the objective is to compare the performance and interpretability of the used methods to derive a classifier that has the "best" trade-off between these dimensions. The code is freely available at `https://github.com/thearkamitra/MLforHC_Project3`.

2 Background

Interpretable machine learning techniques can be grouped by various criteria. For the present paper, the difference between intrinsically interpretable (e.g., Rudin, 2019) and post-hoc explainable methods (e.g., Samek et al., 2020) is crucial, since representatives from each class will be used and contrasted. Whereas the former is restricting the complexity of the used model (e.g., linear models), thereby enabling interpretability, the later uses methods that analyzes the model following the training (Molnar, 2020). Crucially, no consensus has been reached if interpretable or explainable methods are superior (for a partial assessment of posthoc methods, see Samek, 2021). Therefore, this paper compares methods from both classes to add novel insights to the literature, specifically Random Forests (RF), Convolutional Neural Networks (CNN) with SHAP values, Logistic Regression (LR) with L1 regularization, RuleFit, and GradCAM.

First, for Random Forests (Breimann, 2001), feature importance can be calculated as the accumulated impurity decrease within each tree. Second, Convolutional Neural Networks (CNNs) can be augmented with Shap values to explain their predictions. In a nutshell, the prediction in SHAP is explained by assessing each feature's contribution to the final prediction through simulating iteratively that only parts of the feature set are present, while some are discarded (see Lundberg & Lee, 2017 or Molnar, 2020 for a more formal treatment). Third, Logistic regression can be augmented with L1 penalty to perform feature selection (Hastie et al., 2009). The coefficient weights, given equal scaling, can be interpreted as rough feature importance measures. Fourth, RuleFit (Friedman & Popescu, 2008) enhances sparse linear models through interactions that are learned as decision-rules. These rules are generated via random forests or gradient boosting and added to the feature space for a lasso regression. Finally, GradCAM (Selvaraju et al., 2019) generates visual explanations through the use of gradients of any target concept (e.g., tumor) flowing into convolutional layers. This information is leveraged to produce a transparent localization map that emphasizes the relevant regions in the image for the prediction.

The previous paragraph demonstrated the plethora of techniques available for explainable predictions, posing the question of how to properly compare and select methods. Research has defined general desiderata for machine learning explanations (Miller, 2019; Swartout & Moore, 1993). For example, Swartout and Moore (1993) mention five desiderata, including fidelity, understand-ability, sufficiency (i.e, of included knowledge), construction overhead, and efficiency.

3 Methodology

The following section outlines the methodology of the report, starting with a description of the dataset. Then, the analytical strategy will be described.

3.1 Dataset

This project is based on the *Kaggle Brain Tumor Detection dataset* (Kaggle, n.d.). The dataset contains n = 278 brain slices of MRI scans that are labeled as "with tumor" (n_{tumor} = 167) or "without tumor" ($n_{notumor}$ = 111). The data was imported using pre-built data loading scripts, during which images were resized to 128 pixels on the shortest side and center cropped to 128 pixels on the longest side. Next to the images themselves, $n_{radiomics}$ = 474 radiomics features were loaded with a pre-built script (e.g., $original\_shape2D\_Elongation$). There were no missing values present in the dataset. A visual inspection of the tumor and non-tumor images showed no abnormalities, except for a few images having watermarks. There is an unequal distribution among the tumor and non-tumor classes. For the deep learning

frameworks, this has been considered to further improve the results.

3.2 Analytical Strategy

In the following subsection, the analytical strategy for the analysis will be described. First, rationales for model selection will be provided. Second, the approach for training and performance evaluation will be described, including a description of the framework for model comparison and the performance enhancement approach.

3.2.1 Model selection

The baseline models, i.e., CNN and Random forests, were pre-selected, thus, no rationale is provided. For interpretable models both a simple method as well as a more sophisticated one was selected. For the former, a logistic regression with L1 regularization and standardized features was used. L1 regularization prunes some of the features, thereby improving interpretability of the output. Standardization enables a direct comparison of feature coefficients. For the latter, RuleFit has been used as a more sophisticated method that takes into account interaction effects in the form of decision-rules (Friedman & Popescu, 2008). GradCAM is one of the most widely used posthoc methods, which provides visual explanations. One of the objectives was to select methods which are computationally inexpensive, which is why perturbation-based models were not considered. Class Activation Maps (CAM) are one of the simplest methods that can be applied on top to generate the heatmap but since the model has two Dense layers, CAM fails.

3.2.2 Training & evaluation

The analysis has been conducted with the *PyTorch* library (Paszke et al., 2019) for model creation and *scikit-learn* (Pedregosa et al., 2012) for machine learning models and cross-validation. Data was initially split into 80-10-10 training-validation-test data. Due to the low sample size, training and validation data has been merged, resulting in a 90-10 training-test-split. 10-fold Cross-validation was applied to fit the model and produce estimates of the generalization error.

A performance evaluation framework based on two factors is applied to determine the optimal method. Classification accuracy and validation loss have been used as a performance evaluation metric and for saving the best models, respectively. Interpretability is assessed qualitatively by assessing the faithfulness and human interpretability of the methods.

To generate the performance in the aforementioned evaluation framework, each method was trained with specific settings and training strategies. Random Search was used to identify the best parameters of the random forest across *n_estimators* = 1800, *max_features* = 10, *max_depth* = 10, *min_samples_split* = 5, *min_samples_leaf* = 4.

For the binary classification task *modified cross entropy* loss on *softmax* outputs was used as the loss function. The network training was conducted with the *Stochastic Gradient Descent* optimization method. *Early stopping* was used to avoid overfitting with accuracy as stopping criterion. The network architecture was pre-defined and consisted of six convolutional layers with ReLu activation function plus additional layers.

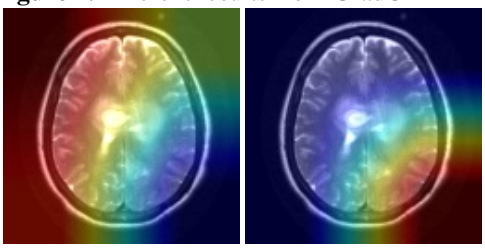
For the (inverse of) regularization strength parameter C, the parameter values [0.001, 0.01, 0.1, 1, 10, 100, 1000] were used in GridSearch cross-validation. Among solvers suitable for L1 penalty, the saga solver was chosen (over *liblinear*) due to the large number of features. Features were standardized to enable ranking of coefficient sizes in order to assess feature importance.

For RuleFit, a Random Forest Classifier was used to determine the decision-rules to ensure compatibility with the baseline. The regularization strength parameter C was determined via crossvalidation, testing the set [0.005, 0.01, 0.1, 1, 10, 100, 1000]. Due to convergence issues, the smallest value was set to 0.005 instead of 0.001. Default values $n\_estimators = 500$ and $max\_depth = 10$ were chosen as hyperparameters, since the identified best hyperparameters for the baseline random forest did not improve test accuracy, but increased the feature space. For the interpretability output, feature importance scores are used. For linear predictors, feature importance is based on standardized predictors. For the rules, the feature importance is based on coefficients and support of the feature in the data, which represents the share of data points to which the specific rule applies (see Molnar, 2020).

GradCAM has been implemented as a different class which takes in a model and selects a layer from the model for registering the forward and backward hooks for manipulating the gradients. The gradients are always with respect to some scalar. The gradient of the predicted class has been considered by default and the heatmap generated from the most likely class is considered. The baseline model that had been provided has been used for obtaining the GradCAM heatmap from the images. The output of the final activation of the layer is of size (2,2), which is too small to provide a useful visual output. For a more specific output, layer 12 which generates the features is used. For higher resolution, the 3rd layer could have been selected but the

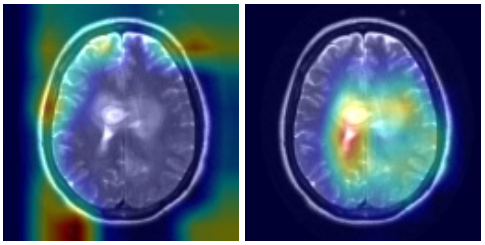
earlier layers are considered to contain very low level information of the images, thus a tradeoff between the resolution and the information from the layer had to be considered.

Figure 1. Different results from GradCAM



(a) Layer 18 with wrong label

(b) Layer 18 with correct label



(c) Layer 12 with wrong label

(d) Layer 12 with correct label

Performance was tuned in several ways. First, for all models, hyperparameter optimization was applied, either via GridSearch or RandomizedSearch (for Random Forests), depending on computational complexity. Furthermore, transfer learning was applied with a pre-trained model. Specifically, a ResNet50 model with pretrained weights was used with optimized hyperparameters. The weights of the last layer have been changed to finetune for the number of classes in the dataset. Transfer learning allows the model to learn more complex representations. Finally, a modified cross entropy loss is used to penalize the model more whenever it misclassifies the class with lower number of samples.

4 Results

The respective classifiers were evaluated on 28 observations of the Kaggle Brain Tumor Detection dataset. The results including the baselines are displayed in Table 1.

The best performing method in terms of raw accuracy performance was Transfer Learning with weights, whereas the least performing method was Linear regression with L1 regularization. For the Random Forest, the feature scores of the different features are extracted and shown in Figure *Figure 4*. We can observe that the feature importance follows an almost exponential graph,

Table 1
Method Results.

Model	**Accuracy %**
Baseline RF	0.714
Baseline CNN + SHAP	0.714
Baseline CNN + GradCAM	0.714
Baseline CNN (weights)	0.785
LR with L1 Regularization	0.679
RuleFit	0.750
Transfer Learning	0.892
Transfer Learning (weights)	**0.964**

where the feature importance drops quite fast. The most important feature is 'original_firstorder_10Percentile'. SHAP values for the CNN predictions of one exemplary tumor and one exemplary non-tumor image are shown in *Figure 3*. The plot shows the original image in the first column, followed by the SHAP values for each class "no tumor", "tumor" in the second and third column, respectively. The CNN correctly identifies the white tumor area as relevant for the tumor class prediction. In the nun-tomor example, it highlights the areas around the brain as relevant for the prediction. Thanks to L1 regularization, only 37 of the 474 features have a weight different from 0 for the Lasso regression. Standardized coefficients for the top 5 features are displayed in *Figure A1* in the appendix. The most important feature is "wavelet-HL_firstorder_Kurtosis". For RuleFit, the feature importance scores are displayed in *Figure A2*, with four of five features in the top five being rules and one feature being a linear feature. The RuleFit resulted in 44 rules and 132 radiomics features. The results of the GradCAM are based on the best weights obtained from task2 on levels 12 and 18. Figure 1a and 1b shows the heatmap obtained from level 18. Due to the very low resolution, the actual features that are important for the model prediction cannot be understood. Figure 1c and 1d shows the heatmap from level 12 and the features are more consistent. Figure 1d shows the heatmap from the correct prediction and as expected, the features which led to the classification are inside the brain while Figure 1c obtained from the other class shows features outside the brain. The interpretibility is quite important for medical applications and our previous work shows that doctors agree with the features that are reported by the model to cause the disease (Mitra et al., 2020).

5 Discussion

There were three main tasks covering two versions of the Kaggle brain tumor dataset, one with radiomics fea-

tures, one with images. The tabular radiomics data enables an assessment of the importance of specific derived features, whereas the image data can be analyzed for important regions of the image for a particular prediction. Three things are noteworthy in regards to performance. First, the results show that transfer learning led to the best tumor classification. Apparently, models like ResNet are able to obtain a higher level understanding of the features of different objects based on their training on a large number of images. Although the model was not trained on medical data, the low-level features obtained from the models also help in the medical domain. The large performance delta compared to other models is thought to be largely attributable to the small sample size that other models were trained on. Second, class imbalance has been mitigated by using modified cross entropy loss, showing better scores for the weighted model versions. Third, among the interpretable models, it is noteworthy that RuleFit outperformed Lasso Regression and the Random Forest. This is in line with expectations, since RuleFit combines both classes by including decision rules learned from a Random Forest in the Lasso regression.

Regarding interpretability, interpretability varied strongly with the used data source. Visual explainers like GradCAM or SHAP provided intuitive explanations, which provided solutions with eye validity. In contrast, the interpretable models provided lists of important features (or decision-rules), but the radiomics features require domain expertise to be informative. RuleFit provides transparency on the most important decision-rules. However, the rules often contain several splits and are very long, making the process of understanding all rules associated with particular features effortful.

Consequently, the most ideal method in regard to performance and explainability would be a combination of transfer-learning and SHAP values as a posthoc method. By using this method combination, no major tradeoff in regard to performance or interpretability needs to be performed.

6 Conclusion

In this assignment, a range of intrinsic and post-hoc methods have been utilized to predict and explain MRI tumor presence. Specifically, a random forest and CNN with Shap values were trained as base models. Further interpretable and explainable models included logistic regression with L1 regularization, the RuleFit method, and GradCAM. A comparison of methods based on a structured comparison framework indicated the ideal method is transfer learning with SHAP values. Future

Figure 3. Tumor and Non-Tumor Example with SHAP-Values

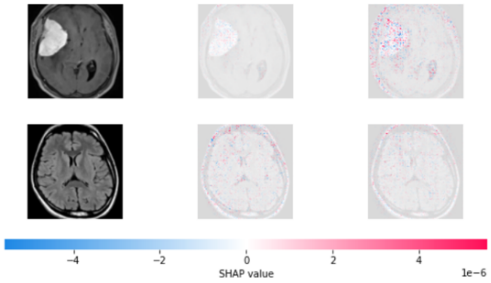
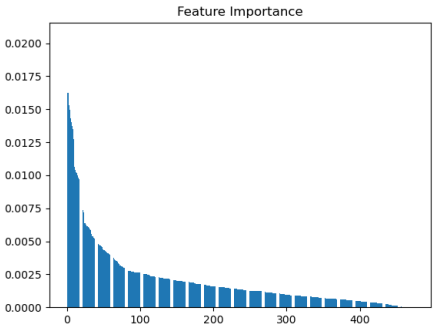


Figure 4. Importance Scores for Random Forest



work could evaluate the methods with real doctors on real tasks to get a more realistic assessment to match the claim being made (Doshi-Velez & Kim, 2017) and assess interpretability in regard to various desiderate in a quantified manner. Conceptual work should refine our understanding of a good explanation and specifically offer feasible ways to assess methods in practice.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop* (pp. 72–78). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-1909 doi: 10.18653/v1/W19-1909

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv. doi: 10.48550/ARXIV.1702.08608

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer. doi: 10.1007/978-0-387-84858-7

Jerome H. Friedman, & Bogdan E. Popescu. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954. doi: 10.1214/07-AOAS148

Kaggle. (n.d.). *Brain tumor detection dataset.* Retrieved from https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. Retrieved from https://www.sciencedirect.com/science/article/pii/S0004370218305988 doi: 10.1016/j.artint.2018.07.007

Mitra A., Chakravarty, A., Ghosh, N., Sarkar, T., Sethuraman, R., & Sheet, D. (2020). A systematic search over deep convolutional neural network architectures for screening chest radiographs. In *2020 42nd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 1225–1228). doi: 10.1109/EMBC44109.2020.9175246

Molnar, C. (2020). *Interpretable machine learning. a guide for making black box models interpretable*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine learning in python.
doi: 10.48550/ARXIV.1201.0490

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. doi: 10.1038/s42256-019-0048-x

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. doi: 10.1109/JPROC.2021.3060483

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359.

Swartout, W. R., & Moore, J. D. (1993). Explanation in second generation expert systems. In J.-M. David, J.-P. Krivine, & R. Simmons (Eds.), *Second generation expert systems* (pp. 543–585). Berlin, Heidelberg: Springer Berlin Heidelberg.

Swiss Federal Statistical Office. (23.04.2021). *The number of mri devices in hospitals has increased by 25% in 5 years.* Retrieved from https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/press-releases.assetdetail.16584132.html

Appendix

Additional plots

The following section contains additional plots:

Figure A1. Top 5 Standardized Coefficients Logistic Regression

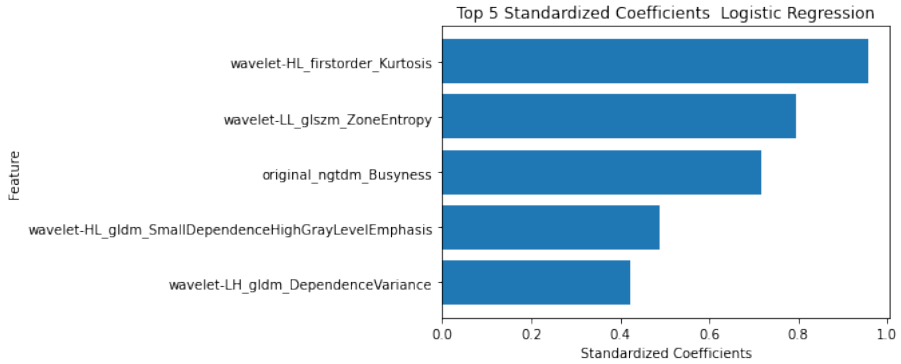


Figure A2. Top 5 Feature Importance Scores RuleFit