Explainable Deep Neural Networks for MRI Based Stroke Analysis

Master of Science in Engineering - Specialisation Project 2 (VT2)

Loran Avci* Beate Sick† Helmut Grabner‡

August 31, 2021

Abstract

Currently, most deep learning models are still black-box methods. Particularly in medical applications, not only the performance but also the explainability of artificial intelligence (XAI) is crucial. In this paper, models are examined which are used to diagnose patients with ischemic strokes. Two different methods of XAI are used to explain deep neural networks: Grad-Cam and Grad-CAM++. They are used to show where the class discriminating pixels are located on a Magnetic Resonance (MR) image. Various experiments show how the explainability of deep learning models can be improved and whether general statements can be made for the classification of stroke patients based on their MRIs. Furthermore, a convolutional neural network (CNN) is implemented to detect an existing stroke and classify the severity of the disability it causes (mRS Outcome). With an accuracy of 94%, images of patients can be classified as stroke or no stroke. The XAI shows that both models can reliably detect brain lesions caused by a stroke. Although the data is unbalanced, the model that predicts the mRS outcome has an overall accuracy of 92%. It is shown that there are differences in the explanations for mRS Outcome 3-6 and mRS Outcome 0-2. By using Grad-CAM, lesions in the brain can be detected, which are even overseen by experienced neurologists. In addition, it is possible to simplify models without significant performance loss by using Grad-CAM. The resulting explanations can thus serve experts such as neurologists and physicians as a basis for new hypotheses or even help them to improve their diagnostic quality.

*Zurich University of Applied Sciences (ZHAW), avci@zhaw.ch - Author
†University of Zurich (UZH), Zurich University of Applied Sciences (ZHAW) - Supervisor
‡Zurich University of Applied Sciences (ZHAW) - Supervisor

Contents

1 Introduction

An ischemic stroke occurs when there is a sudden blockage of the cerebral arteries. This type of stroke accounts for over 80% of all strokes (Feign et al., 2003). The cause of an ischemic stroke is often an Emboli, e.g., a clogged blood vessel. In this case, a stroke is treated by reopening the vessel by dissolving the thrombus with medication (Kraft et al., 2012). Early detection of an ischemic stroke is crucial. During a typical acute ischemic stroke, 120 million neurons are destroyed per hour. Compared to a normal brain, an ischemic brain ages 3.6 years per hour unless treated (Saver et al., 2015). To diagnose stroke patients, experts rely on clinical patient data and magnetic resonance imaging (MRI). MRI can be used for the early detection of ischemic strokes and to discriminate brain tumors and other valuable applications in brain research (Chilla et al., 2015).

The success of machine learning (ML) algorithms in image recognition in recent years is also finding its way into the application of medical image analysis. Using Deep Learning (DL) to be more precise Deep Convolutional Neural Networks (CNN), it is possible to detect hierarchical relationships in the data without extensive feature engineering (Ker et al., 2018). Nevertheless, artificial neural networks are still regarded as black-box methods. The complex correlations that the neural network determines are neither explainable for domain experts nor technical users, let alone interpretable. However, these ML algorithms are increasingly relevant for predictive processes in critical decisions (Alber et al., 2019). Many methods and approaches have been developed and introduced in recent times, which should shed light on the black boxes. These methods can be summarized under the keyword Explainable Artificial Intelligence (XAI).

1.1 Objectives

In this paper, we want to implement a post-model XAI method. MRIs and clinical patient data in tabular form from stroke subjects and a control group serve as the data basis. The XAI method should be applied to a pre-trained deep learning model, e.g., a CNN, classifying brain MRIs of stroke or TIA patients. The XAI method should be able to account for the individual predictions of the pre-trained deep learning model by highlighting the relevant pixels of an image. In order to achieve this objective, the following steps are taken:

- Establish a definition for XAI.
- Investigation and evaluation of different XAI methods and selection of a method.
- Determination of optimal preprocessing steps for the selected method.
- Investigation of the pre-trained model for parameter determination of the XAI method.
- Interpretation of the heatmaps resulting from the XAI method with experts.

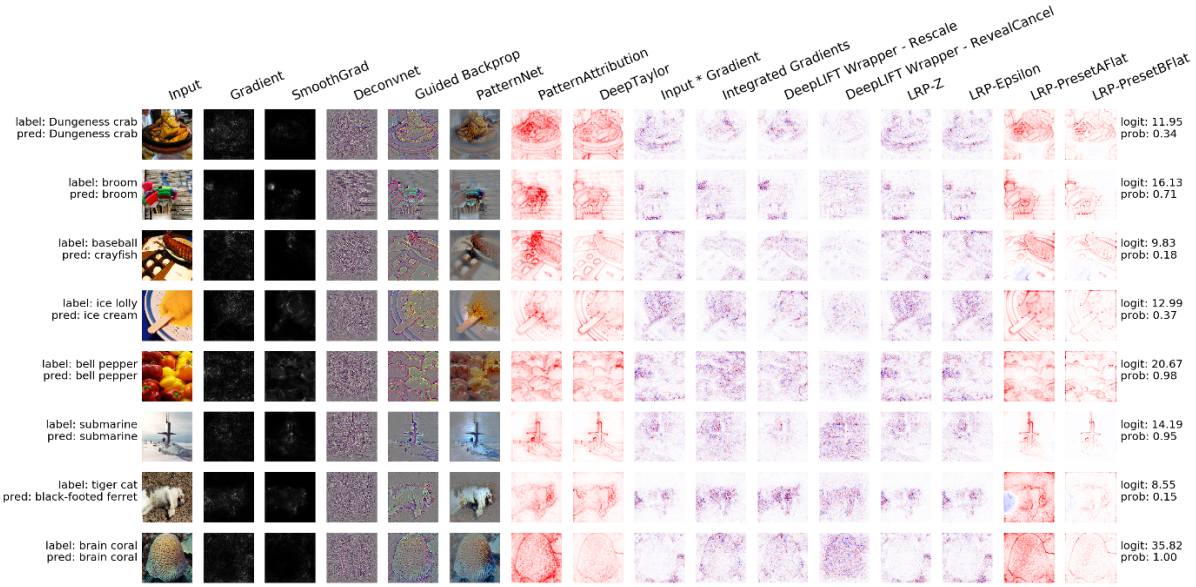

Figure 1: Overview of different XAI methods implemented in the Python library iNNvestigate (Alber et al., 2019). The methods which are shown in the figure mostly show the location of edges on the input image.

1.2 Related Work

Classification of MRIs showing ischemic stroke is possible with over 95% accuracy, as shown in the paper by Herzog et al. (2020). This paper also addresses the uncertainty in prognosis that is often left out of deep neural networks. In order to provide an uncertainty measure for each image-level prediction, the authors implement a Bayesian CNN model architecture. A degree of uncertainty for a prediction is a first step towards relieving artificial neural networks of the image of black boxes. The methods that have been developed and used to analyze artificial neural networks are numerous and often go by the keywords interpretability or explainability (David et al., 2010, Smilkov et al., 2017, Zeiler et al., 2014, Springenberg et al., 2015, Montavon et al., 2017, Kindermans et al., 2018, Sundararajan et al., 2017, Shrikumar et al., 2019, Bach et al., 2015, Ribeiro et al., 2016). Most of these methods work by highlighting areas of the input, such as pixels of an image, that were important in classifying an image (Leavitt and Morcos, 2020).

After evaluating different XAI methods (cf. Figure 1), it is noticeable that many of the methods often only detect edges. While this can be useful, it does not correspond to our idea of an explanation for a neural network. Class Activation Mapping (CAM) is an alternative approach to the methods shown in Figure 1, but this method requires a global average pooling layer (GAP) in the CNN, limiting this method's usage (Zhou et al., 2016). CAM and its further developments, namely Grad-CAM and Grad-CAM++, generate heat maps by multiplying each feature map of the last convolutional layer by the associated weight of the predicted class (Zhou et al., 2015, Selvaraju et al., 2019, Chattopadhay, et al., 2018). Grad-CAM and Grad-CAM++, do not require any adaptations to the model architecture, which leads to increased flexibility (Selvaraju et al., 2019, Chattopadhay et al., 2018). Ac schematic visualization of the algorithms is shown in Figure 5. The Grad-CAM and Grad-CAM++ algorithms are explained in more detail in chapter 3.3 respectively 3.4.

The literature has shown that these XAI methods can deliver promising results when applied in the medical imaging analysis domain (Zhang et al., 2021, Fernández et al., 2020). Zhang et al. (2021) show that Grad-Cam can help interpret deep learning models. The authors classify different types of multiple sclerosis based on MRIs of the brain. According to the authors, the Grad-Cam algorithm outperforms similar methods regarding the interpretability of the models. They also show how crucial the choice of model architecture is for classification. The fact that XAI methods can also be used for 3-dimensional MRIs is shown by the study of Kan et al. (2020). In their work, the authors interpret 3D CNN for the classification of brain MRIs. They search for gender-specific differences in brain activity in healthy male and female subjects. The methods used in the paper include Grad-CAM and guided backpropagation. Further findings from previous studies show how important the choice of the evaluation layer is for XAI methods such as the Grad-CAM algorithm (Pereira et al., 2018). The authors of this paper classify the severity or types of brain tumors based on brain MRIs using CNNs. They compare the heat maps resulting from Grad-CAM for different layers and thus draw attention to this essential factor. There is also justified criticism of these methods as they often generate the same explanations for different and sometimes unreasonable outputs (Rudin, 2019). Another criticism is the lack of falsifiability of these methods, leading to misleading conclusions (Leavitt and Morcos, 2020).

2 Explainable Artificial Intelligence

Classical statistical models such as linear regression models have parameters that can be determined analytically or numerically. These parameters can be easily interpreted by those who use these models, making such models very transparent even with many parameters. However, deep learning models do not provide any information on how they arrive at a prediction due to their hierarchical non-linearity. In order to interpret or explain Deep Learning models, one must first define terminology for XAI for these types of models.

The trustworthiness indicates how certain a model is in a prediction. A measure for certainty is achieved, for example, with the help of confidence intervals. In the case of individual predictions, it also helps to specify the probabilities for the individual classes in the case of classifications. Interpretability is defined as being able to understand why a system has made a decision. An example of this is what variables led an ML algorithm to classify a patient in a certain way. Explainability, on the other hand, corresponds to tailored interpretability, i.e., a user-dependent interpretation. This is justified because an interpretation requires different explanations for different recipients (Van der Schaar, 2020). In applying ML or DL models in the medical field, an explanation for a researcher, for example, corresponds to a data-induced hypothesis, while a physician's explanation contains information on which treatment should be recommended to a patient (cf. Figure 1). Thus an explanation has different facets.

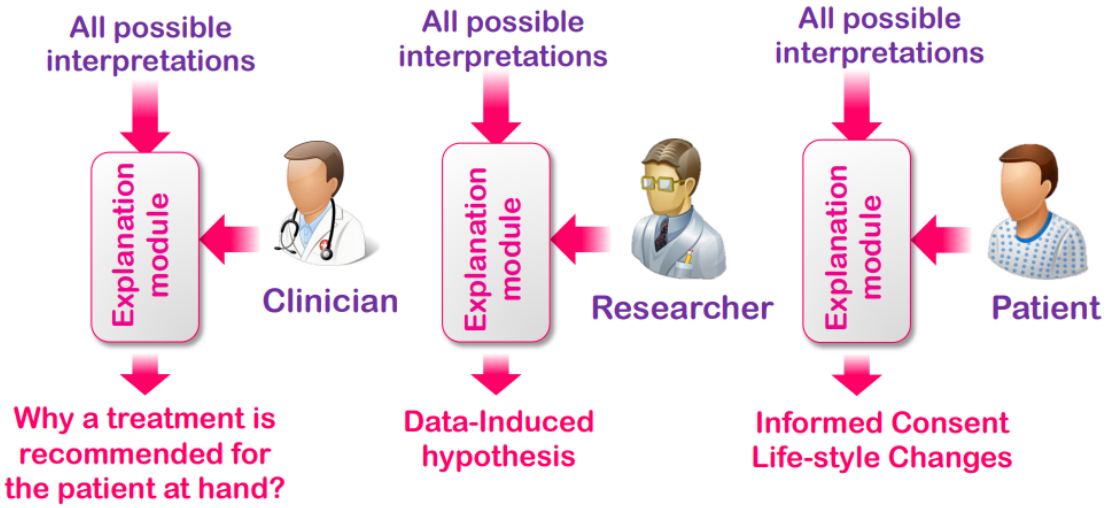

Figure 2: Explanations are dependent on the recipient. Different recipients require explanations that provide different information (Van der Schaar, 2020).

2.1 Recipients

Different recipients require different explanations that have different information content and convey different levels of detail. For example, in the case of classification of medical images using Deep Learning, an explanation for an AI user will highlight the areas of the image that are important for the prediction of a model. On the other hand, research institutions may not be interested in explanations for individual predictions but in global or aggregate explanations or patterns that the model has learned. Such insights could then lead to hypotheses that could be validated with new data (Samek and Müller, 2019).

2.2 Information Content

Different types of explanations also provide the recipient with different insights into a model. Roughly speaking, four different types of information content can be distinguished concerning AI.

- Explaining learned representations: This type explains by representing what the model has learned about complex abstractions, e.g., what the model has learned about the category "cat" by generating a typical image for the category (Bau et al., 2020).
- Explaining individual predictions: This type highlights the features of individual predictions that led to the classification. This could be reflected in medical image analysis by visualizing a heat map for the class-discriminating pixels (Zhou et al., 2015).
- Explaining model behavior: This type goes further than explaining individual predictions towards a more generalizable understanding of system behavior. For example, heat maps of individual predictions are clustered to identify strategies for

predicting classes (Lapushkin et al., 2019).

- Explaining with representative examples: This type explains by identifying representative training examples. These can be, for example, training images that have a high impact on the prediction of test images. This can also be used to detect model biases or train models more robustly (Koh and Liang, 2017).

2.3 Role

In addition to the recipient and the information content, the purpose of an explanation must also be considered. Explanations are relative, so it makes a difference whether the purpose is to show how a model arrived at a specific prediction or whether the recipient is interested in how the explanation compares to an alternative. Two aspects need to be clarified. On the one hand, it must be understood what the intention of the XAI method is (for example, what does a CAM show). On the other hand, the user's intention of the XAI method must be clarified, i.e., what the explanation should be used for (Samek and Müller, 2019).

3 Material and Methods

3.1 MRI Image Data

For this study, data from the neurological department of the University Hospital Zurich was used. The data collected and processed can be divided into two types of patients. The first group consists of those who had suffered an ischemic stroke. The second group consists of patients who had suffered a transient ischemic attack (TIA). TIA patients suffer from similar neurological damage as patients with an ischemic stroke, but the damage usually disappears after 24 hours without leaving any visible lesions on the MRI (Herzog et al., 2020). In total, MRIs of 511 patients from one of these patient groups are available. The imaging technology used to acquire the MRIs is diffusion-weighted imaging (DWI). DWI also detects marginal changes in water diffusion that occur in an ischemic brain. The method is superior to other MRI imaging technologies, such as T2-weighted MRI, detecting acute stroke (Lutsep et al., 2004). The MRIs are from the axial plane, which divides the human body into superior and inferior, i.e., head and feet (cf. Figure 3).

An experienced neurologist labels the image data for a visible lesion to determine the ground truth. Thus, there is a "Stroke" or "no Stroke" label for each stroke and TIA patient at the image level. Each patient can be allocated an average of 30 images, with a minimum of 21 and a maximum of 46 images available (Herzog et al., 2020). On average, 12.5 images per patient show a lesion on the MRI image. The 511 patients are divided into 355 patients with an ischemic stroke and 156 TIA patients. Approximately 30% of the images of stroke patients show a visible lesion on the MRI. Of the TIA patients, no stroke is visible on the MRIs on 100% of the images. The image data used for this study has the dimension 192x192x1 pixel. Images that are entirely black and therefore contain no information are removed from the data set. Furthermore, the images are normalized so that each image has a mean of zero and a variance of one. Another processing step

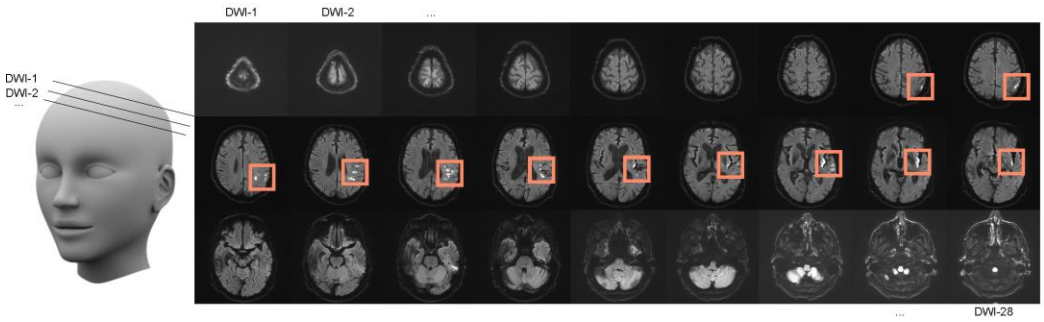

Figure 3: MRI for a patient with ischaemic stroke. The individual pictures show a two-dimensional image from an axial perspective. Strokes are usually visible on the MRI as bright anomalies (Herzog et al., 2020).

is cropping and interpolation. The contour of the brain is determined for each image using an image-dependent threshold. The image is then resized to 192x192x1 using the determined contours and bicubic interpolation.

3.2 Tabular Patient Data

In addition to the image data, further clinical patient data is available in tabular form. In total, the tabular data contains 23 different attributes and 497 observations, where each observation can be assigned to a patient using a patient ID. The data contains demographic information such as age and gender and clinical information such as the presence of a stroke or TIA, previous illness such as diabetes, or the degree of disability three months after the onset of a stroke. The data can be divided demographically into 309 female patients and 188 male patients. The age of the patients can be approximately described as a left-skewed unimodal distribution with a median age of 71. The data contain information from 315 stroke patients and 182 TIA patients.

In addition to the patient ID, only the variable "mrs_3months" is included in this work. This variable contains a patient's modified Rankin Scale (mRS) outcome after a stroke. The mRS outcome is an ordinal variable that indicates the degree of disability after a stroke. The mRS outcome ranges from zero, which involves no symptoms, to the mRS outcome of six, which involves death due to the stroke (van Swieten et al., 1988). The severity of disability between the two extremes, zero and six, ranges from light to severe (cf. Table 1). The distribution of the mRS outcome in this data set is highly unbalanced and zero-inflated (cf. Figure 4).

Table 1: The modified Rankin Scale (van Swieten et. al., 1988). The severity of a stroke increases with an increasing mRS Score.

Score	Description
0	No symptoms at all
1	No significant disability despite symptoms: able to carr out all usual duties and activities
2	Slight disability: unable to carry out all previous activities but able to look after own affairs without assistance
3	Moderate disability: requiring some help, but able to walk without assistance
4	Moderately severe disability: unable to walk without assistance, and unable to attend to own bodily needs without assistance
5	Severe disability: bedridden, incontinent, and requiring constant nursing care and attention
6	Dead

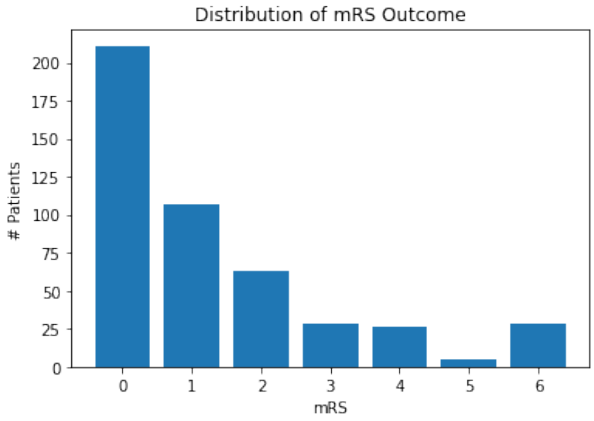

Figure 4: The distribution of mRS outcomes after three months is highly unbalanced. Most Patients are scored with an mRS outcome of zero, while only very few have an mRS outcome higher than two.

3.3 Crad-CAM

Selvaraju et al. (2019) show that the Grad-CAM algorithm uses the gradient information that flows into the final convolutional layer to understand the relevance of each neuron of the CNN to a specific decision. Although this is not their intention, the algorithm is not limited to the final convolutional layer. The algorithm can visualize any non-one-dimensional layer. To create a class-discriminating heat map $L^c_{Grad-CAM}$, the gradient for the score y^c of a class c is first calculated for a feature map A^k of a convolutional layer, i.e., $\frac{\delta y^c}{\delta A^k}$. This must be calculated without evaluating the softmax layer.

This calculated gradient is then passed back and averaged-pooled to obtain the relevance weights of the individual neurons w^c_k:

$$w^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\delta y^c}{\delta A^k} \quad (1)$$

where Z denotes the total number of elements in a feature map. The weights w^c_k reflect the importance of a feature map k for a target class c. By a weighted linear combination followed by a ReLU function, we get:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k w^c_k A^k\right) \quad (2)$$

The ReLU function is used to display only features that have a positive influence on the chosen class. Without applying this ReLU function, the Grad-CAM would highlight regions that are not only assigned to the desired class and would therefore localize poorer. The Grad-CAM algorithm results in a heatmap with the same size and dimension as the layer on which Grad-CAM is applied. For example, if the size of the evaluation layer of the CNN for Grad-CAM is 12x12 and the input image is 192x192, the heatmap can be resized to the input image size to overlay the two images (Selvaraju et al., 2019).

3.4 Grad-CAM++

The authors of Grad-CAM++ extend the Grad-CAM algorithm by applying the exponential function to the classification score y^c and thus obtain Y^c. The weights w_k^c are further extended by the second and third derivatives of the gradient of the classification score:

$$w_k^c = \sum_i \sum_j \left[\frac{\frac{\delta^2 Y^c}{(\delta A_{ij}^k)^2}}{2\frac{\delta^2 Y^c}{(\delta A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \left\{ \frac{\delta^3 Y^c}{(\delta A_{ij}^k)^3} \right\}} \right] \cdot ReLU\left(\frac{\delta Y^c}{\delta A_{ij}^k} \right) \quad (3)$$

The iterators (i, j) and (a, b) apply to the same activation map A^k and are used to avoid any confusion. The heat map for Grad-CAM++ is then formed by a linear combination of the weights and the activation maps, to which a ReLU function is applied:

$$L_{ij}^c = ReLU\left(\sum_k w_k^c A_{ij}^k \right) \quad (4)$$

The authors of Grad-CAM++ justify their algorithm with a better performance in the localization ability, especially in the case of multiple occurrences of the same class on an image (Chattopadhay et al., 2018).

3.5 CNN Achitectures

All models used in this work are based on convolutional networks (LeCun, 1989). This type of deep learning model has proven to be highly successful for application to image data (Goodfellow et al., 2016). In this work, the architecture developed by Herzog et al., 2020 serves as the baseline model. This architecture leans heavily on the architecture

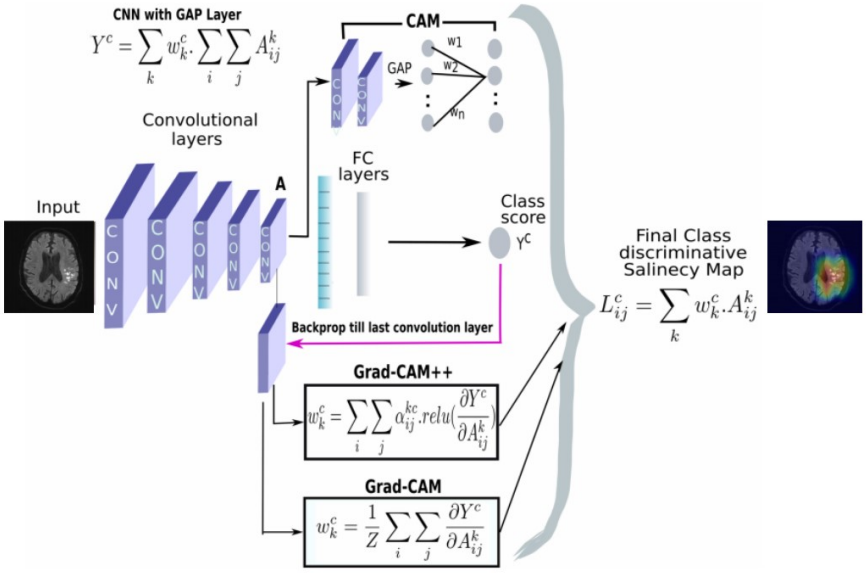

Figure 5: Overview of different CAM methods (Chattopadhyay et al., 2018). To obtain a weight matrix, the classification score is backpropagated to the last convolutional layer.

of VGG (Simonyan & Zisserman, 2014). All models were implemented in Keras with Tensorflow 1.x respectively 2.x backend (Chollet & others, 2015). The code is made available at Github for reproducibility (https://github.com/avciidp/explainable_ai).

3.5.1 Binary Prediction on Image Level

This model aims to detect whether or not a stroke is visible on an MRI image. The model consists of 16 blocks which all have a similar structure. A block usually consists of a 2D convolutional layer with a filter size of 3x3, a batch normalization layer, a ReLU activation, an MC dropout layer with a dropout level of 0.3, and a max-pooling layer with a pooling size of 2x2. The number of filters in the convolutional blocks increases from 32 to a maximum of 512 filters. A flattening layer does the transition from the convolutional part to the fully connected part. In the dense part of the model, two blocks are used, each consisting of a dense layer, a batch normalization layer, a ReLU activation function, and an MC dropout layer with a dropout level of 0.3. The dense layers have 400 and 100 neurons. This results in a model that has over 16.7 million parameters. The negative log-likelihood respectively the binary cross-entropy is used as the loss function to fit the model (Dürr et al., 2020):

$$Loss = -\frac{1}{n}\left(\sum_{j=1}^{n}\left(\ y_i\ \cdot\ log\Big(\ p_1\Big(\ x_i\ \Big)\Big) + \Big(\ 1 - y_i\ \Big)\ \cdot\ log\Big(\ 1 - p_1\Big(\ x_i\ \Big)\Big)\right)\right) \quad (5)$$

Where n corresponds to the number of samples and $p_1(x_i)$ corresponds to the predicted probability that the image x_i corresponds to the class ($y_i = 1$) and thus a stroke is visible on the image. Accordingly, $(1 - p_1(x_i))$ is the predicted probability that the image x_i corresponds to the class ($y_i = 0$), i.e. that no stroke is visible on the image. The Image Data Generator implemented by Keras is used for data augmentation to avoid overfitting, and early stopping is applied based on the lowest loss value in the validation set. Based on experiments, the model architecture for binary classification will be subsequently modified several times.

3.5.2 Ordinal Prediction on Image Level

The purpose of this model is to classify the mRS outcome for a single MRI image. As mentioned in chapter 3.2, in addition to the labels “Stroke” and “no Stroke”, further patient information is available in tabular form. For 497 patients, the ordinal factor variable “mRS_3months” is available with a range of 0-6. The problem here is that these labels are only available at the patient level and not at the image level. Nevertheless, there is an interest in predicting the mRS outcome based on a 2-dimensional image. For this purpose, a labeling method is used in this work, which compromises complexity and feasibility. The ordinal mRS outcomes are first binarised. This means we distinguish between mRS outcome 0-2, which stands for no disability up to mild disability, and mRS outcome 3-6, which leads from moderate disability up to death. Each patient can thus be assigned one of the two mRS outcome labels. In order to apply the labels from the patient level to the image level, a 3-class problem is created. Depending on whether a stroke is

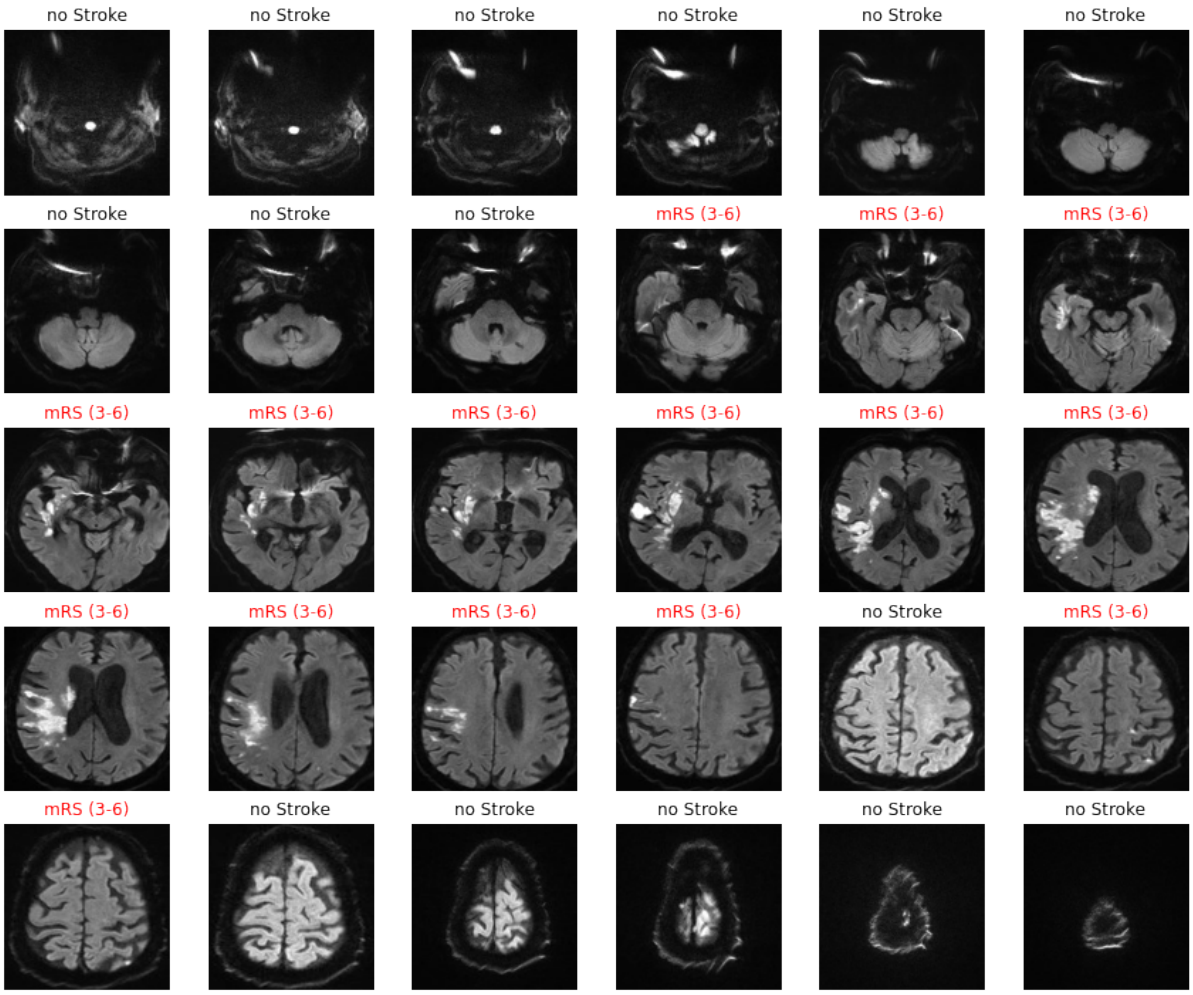

Figure 6: The figure shows how MRIs are labeled for ordinal classification. All images with a stroke label receive the corresponding binary mRS label of the patient.

visible on an MRI image, each patient is assigned the mRS outcome label corresponding to that patient instead of a "stroke" label. MRI images with a "no Stroke" label are not assigned a new label. The following example describes this method: If a patient has an mRS outcome of 5, the mRS outcome label is mRS (3-6). If a stroke is visible on 15 of 30 MRI images in this patient, these 15 MRI images receive the label mRS (3-6). This means that this patient has 15 images with the label "no Stroke" and 15 with mRS (3-6). Graphical visualization of this example can be found in Figure 6.

The architecture for classifying the mRS outcome corresponds to a simplified form of the binary classification model. The model has five convolutional blocks, all of which have the same structure. The blocks consist of a 2-D convolutional layer with kernel size 3x3, a batch normalization layer, a ReLU activation, an average pooling layer with pooling size 2x2, and a dropout layer with a dropout level of 0.3. The first convolutional layer has 32 filters and doubles in each convolutional layer. The transition from the convolutional part to the flat part of the model is done with a GAP layer followed by a batch normalization layer, a ReLU activation, and a dropout layer. This results in a model with less than 1.6 million parameters. A visualization of the model is shown in the appendix. How this architecture is achieved will be explained in more detail in chapter 4.1.4. Due to the

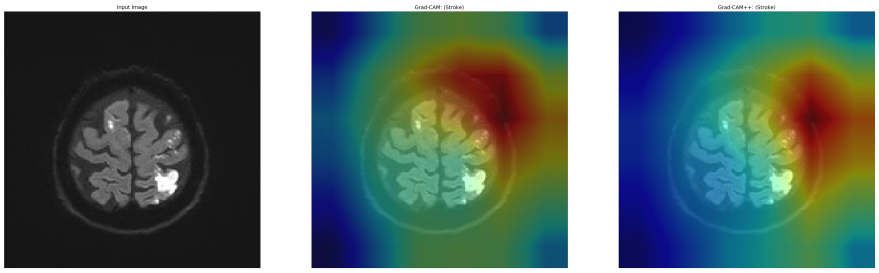

Figure 7: The input image (left) is correctly classified as a stroke by the model. Both the Grad-Cam (cemter) and Grad-CAM++ (right) algorithms do not highlight the lesion.

multiclass problem, the categorical cross-entropy is used as a loss function to fit the model (Dürr et al., 2020):

$$Loss = -\frac{1}{n}\left(\sum_{j \ with \ y_j=0} log(p_0(x_j)) + \sum_{j \ with \ y_j=1} log(p_1(x_j)) + \sum_{j \ with \ y_j=2} log(p_2(x_j)) \right) \quad (6)$$

Where n stands for the number of samples j stands for the different classes ($y_j = 0$) = "no Stroke", ($y_j = 1$) = mRS (0-2) and ($y_j = 2$) = mRS (3-6). To avoid overfitting, the ImageDataGenerator implemented by Keras is used for data augmentation and earlystopping is applied based on the lowest loss value in the validation set.

4 Results and Experiments

Herzog et al. (2020), which implemented the baseline model, showed that MC dropout could significantly improve the model. The test accuracy for the prediction at the image level was 95.52% [95.18%, 95.83%] for the best model. Further performance measures and uncertainty measures can be found in Herzog et al. (2020).

4.1 Stroke vs. No Stroke

Using the Grad-CAM and Grad-CAM++ algorithms, we will show which areas in the image led to the prediction of the desired class. The former results show that the class discriminating pixels are distributed over the whole area of the brain (cf. Figure 7). Even areas of the MRI where no brain is visible are highlighted. Regions on which a lesion is visible, i.e., bright spots on the MRI that indicate a stroke, are touched but only inaccurately detected. It is also noticeable that the heat maps often have high activation on the same side of the brain, indicating a bias in the Grad-CAM implementation. These results are not very helpful since they do not explain anything and therefore cannot be interpreted. Therefore, in this paper, we experimentally investigate how to improve the explanatory power of the neural network.

A mirroring experiment is carried out in which the same input image is evaluated twice by the two algorithms, once in original form and once mirrored on the vertical axis. Furthermore, it is investigated whether cropping the brain improves in terms of activations

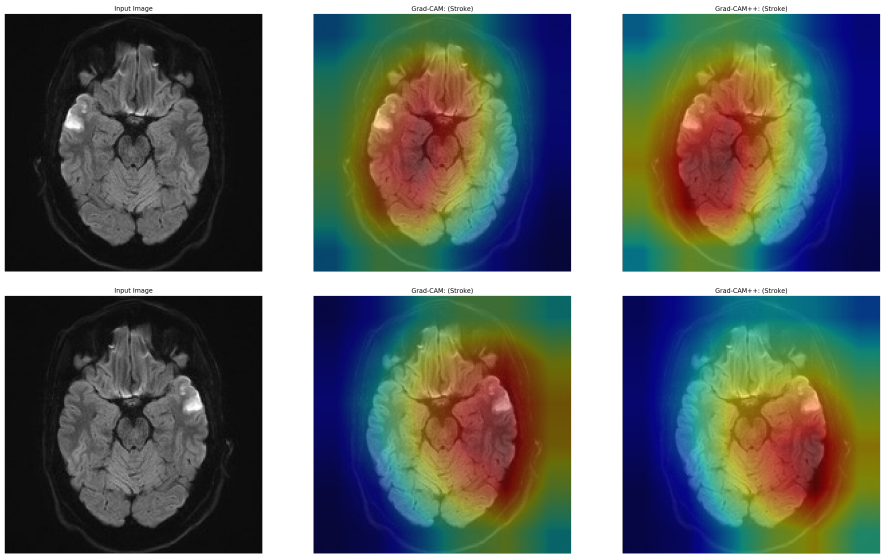

Figure 8: The experiment shows that with a mirrored image, the explanations are also approximately mirrored.

outside the brain. Finally, a layer iteration experiment is performed to investigate the influence of the evaluation layer on the resulting heat map.

4.1.1 Mirroring

Observable behavior is that the Grad-CAMs often highlight similar regions frequently located on the same side. Therefore, we want to investigate whether there is a miss implementation of the Grad-CAM algorithm or whether a bias in the model could be the reason for this. This question is to be clarified with the help of a mirroring experiment. For this purpose, an MRI image showing an evident lesion is passed to the baseline model for prediction, and then the Grad-CAM and Grad-CAM++ for the class “Stroke” are visualized. The same procedure is repeated with the same input, but the input is mirrored on the vertical axis. Therefore, one would expect that the heat maps would also be approximately mirrored on the vertical axis. On the other hand, if the class-discriminating pixels on the mirrored input were on the same side as on the original input, this would indicate the problems mentioned above. The experiment has shown that the desired effect of mirroring occurs. Both Grad-CAM and Grad-CAM++ show an approximate mirrored heatmap for the mirrored input (cf. Figure 8). Thus, no incorrect implementation of the Grad-CAM algorithms is responsible for insufficient explainability.

4.1.2 Image Cropping

Another feature that interferes with the explanation of the Grad-CAM algorithms is the regions that are highlighted outside the brain. This may be due to the different sizes of the axial slices of the brain (cf. Figure 3). The MRIs that are close to the skullcap visualize only a tiny area of the brain. MRIs that image deeper layers, on the other hand, fill more pixels with relevant information. All images should be cropped to counteract this effect so that the brain covers approximately the same surface area on all MRIs. To

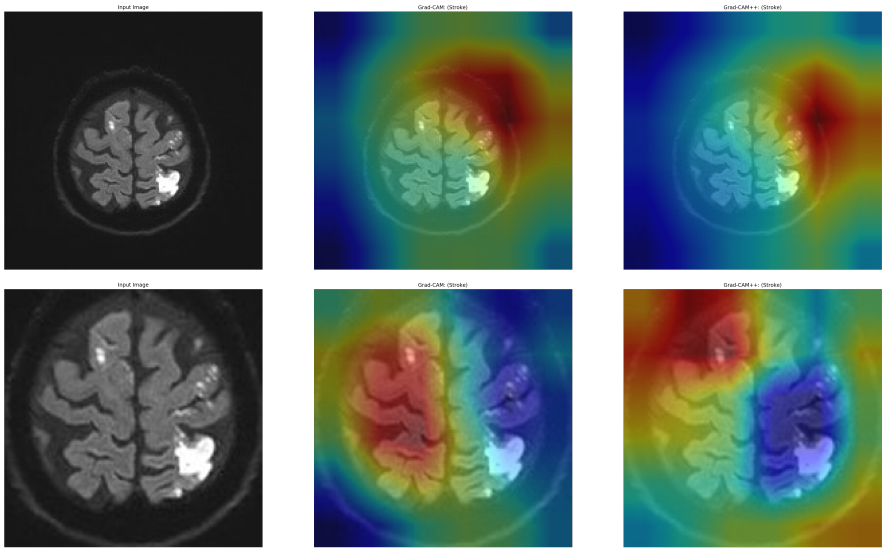

Figure 9: Cropping images leads to less activation outside of the brain in the heat maps.

achieve this, a binary threshold is determined based on each image. This threshold is then used to detect the contours algorithmically. The image is then cropped based on the extrema of the contours. To ensure that the dimensional structure of the images is still the same, the images are scaled to their original size using bicubic interpolation (OpenCV, 2015). The results of the cropping experiments show that Grad-CAM and Grad-CAM++ highlight significantly fewer regions outside the brain. This is to be expected, as the cutting out of the brain also results in fewer informationless pixels that can be marked as class-discriminating in the first place. However, it is not only the regions outside the brain that reveal a change in the heat maps. The cropped MRIs show explanations that are not visible on the uncropped image (cf. Figure 9). Consequently, the results are a step towards better explainability. Image cropping is therefore introduced as a preprocessing step for all remaining results.

4.1.3 Evaluation Layer Iteration

References using XAI on brain MRIs indicate how the choice of model architecture significantly affects the resulting heat maps from Grad-CAM (Zhang et al., 2021). The choice of the evaluation layer for the Grad-CAM algorithm also plays a crucial role in visualizing the heat maps (Pereira et al., 2018). Therefore this experiment will illustrate the influence of the evaluation layer for the baseline model. The baseline model has 81 layers, 70 of which can be used as evaluation layers. In the experiment, an MRI image showing a visible lesion is evaluated by the model, and then a Grad-CAM is created for each evaluable layer. For simplification purposes, only the convolutional layers are visualized in Figure 10. The resulting heatmaps show that after the first five convolutional layers, the model considers completely different features as class discriminating than in the layers before. Especially in the last three convolutional layers, the class discriminating pixels seem to blur. This leads us to assume that the model is too deep and therefore too complex for this problem. The explanations interpretable to an expert are the bright spots that correspond to the lesions. These are already recognized in the first five convolutional layers. With these insights, the model is to be debugged, and it is to be investigated how

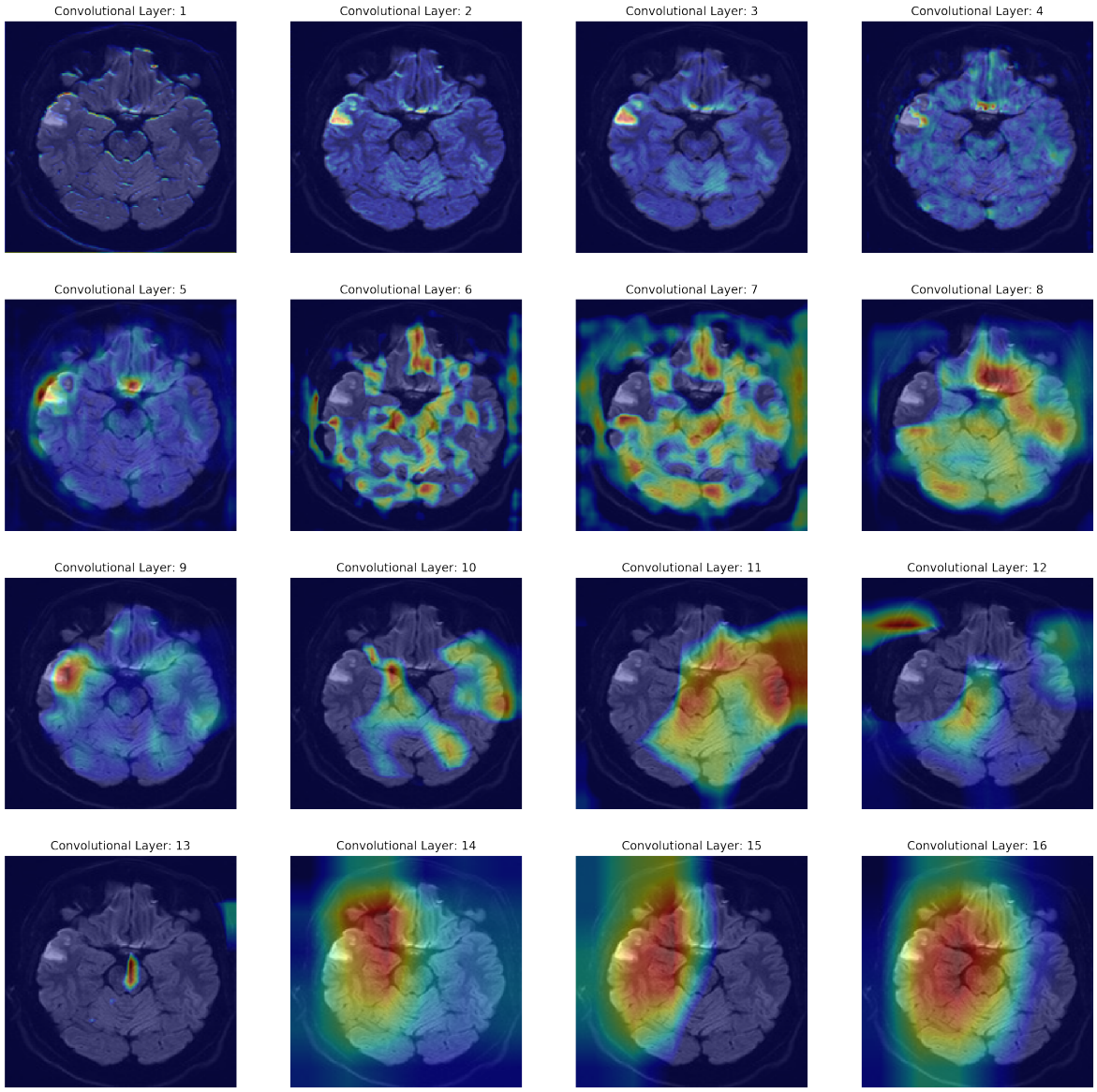

Figure 10: We observe that the class-discriminating pixels that detect the strokes are already recognized after five convolutions by iterating the evaluation layer. These detected features become blurred with increasing complexity.

model performance and explainability are compatible.

4.1.4 Model Debugging

In order to evaluate the influence of the model architecture and model complexity on performance and explainability, different models are implemented. The identical random seed is set for all models, and the same training, validation, and test data is used. The influence of dense layers in the model and differences between flatten and GAP layers are mainly investigated.

The baseline model has an accuracy of 95.52% [95.18%, 95.83%] and has 16.7 million parameters distributed over 16 convolutions and two dense layers (the softmax layer will

not be counted as a dense layer). The transition from the convolutional part to the fully connected part is made using a flatten layer. While the performance is excellent, the XAI visualized by the Grad-CAM is considered qualitatively poor. Therefore, different architectures are trained, and their performance and explainability are evaluated (cf. Table 2).

The "Dense-Model" has an accuracy of 94.83% [93.99% 95.57%] and has 6.3 million parameters distributed over five convolutions and two dense layers. The transition from the convolutional part to the fully connected part is made through a flatten layer. The model performs worse than the baseline model. The output of the Grad-CAM is considered qualitatively poor. The "Flat-Model" has an accuracy of 92.69% [91.71% 93.57%] and has 1.7 million parameters distributed over five convolutions and 0 dense layers. The transition from the convolutional part to the fully connected part is made through a flatten layer. The model performs worse than the baseline model. The output of the Grad-CAM is considered qualitatively poor. The "GAP-Model" has an accuracy of 94.34% [93.46% 95.11%] and has 1.5 million parameters distributed over five convolutional layers and 0 dense layers. The transition from the convolutional part to the fully connected part is made through a GAP layer. The model performs slightly worse than the baseline model. The output of the Grad-CAM is considered qualitatively good. The "Shallow-Model" has an accuracy of 89.77% [88.64% 90.79%] and has 94'000 parameters distributed over three convolutional layers and 0 dense layers. The transition from the convolutional part to the fully connected part is made through a GAP layer. The model performs significantly worse than the baseline model. The output of the Grad-CAM is considered qualitatively very good.

Other performance metrics can be found in Table 2. The table shows that the "Dense-Model" usually performs best. The confidence intervals of the "GAP-Model" and the "Dense-Model" overlap in accuracy, sensitivity, and specificity. Compared to the "Dense-Model", the explainability of the "GAP-Model" performs better. An evaluation layer iteration experiment was conducted for all models. To summarise, it can be said that flatten layer, and dense layer have a suboptimal effect on the explainability of a deep learning model, while a GAP layer instead favors explainability. Due to the excellent performance and the good explainability, the "GAP-Model" is considered the most suitable model. For this reason, the architecture of the "GAP-Model" is used for all further evaluations. Based on the results of the experiments, the activation layer after the last convolution layer is used as the evaluation layer for all other results.

4.1.5 Results

The GAP model presented earlier will be referred to as the Stroke model from now on. Based on the findings of the experiments and the model debugging, another attempt can

Table 2: Overview of performance metrics for the model debugging experiment. All confidence intervals are computed with a Wilson proportion interval.

Architecture	Accuracy [95% Conf.]	Sensitivity [95% Conf.]	Specificity [95% Conf.]	AUC	Negative Log-Likelihood
Dense-Model	**0.9483 [0.9399 0.9557]**	**0.7888 [0.7562 0.8182]**	0.9928 [0.9886 0.9955]	**0.8908**	**0.1668**
Flat-Model	0.9269 [0.9171 0.9357]	0.6923 [0.6562 0.7262]	0.9924 [0.9881 0.9952]	0.8424	0.3871
GAP-Model	0.9434 [0.9346 0.9511]	0.7572 [0.7231 0.7883]	0.9954 [0.9917 0.9974]	0.8763	0.1994
Shallow-Model	0.8977 [0.8864 0.9079]	0.54 [0.5019 0.5776]	**0.9975 [0.9945 0.9988]**	0.7687	0.2821

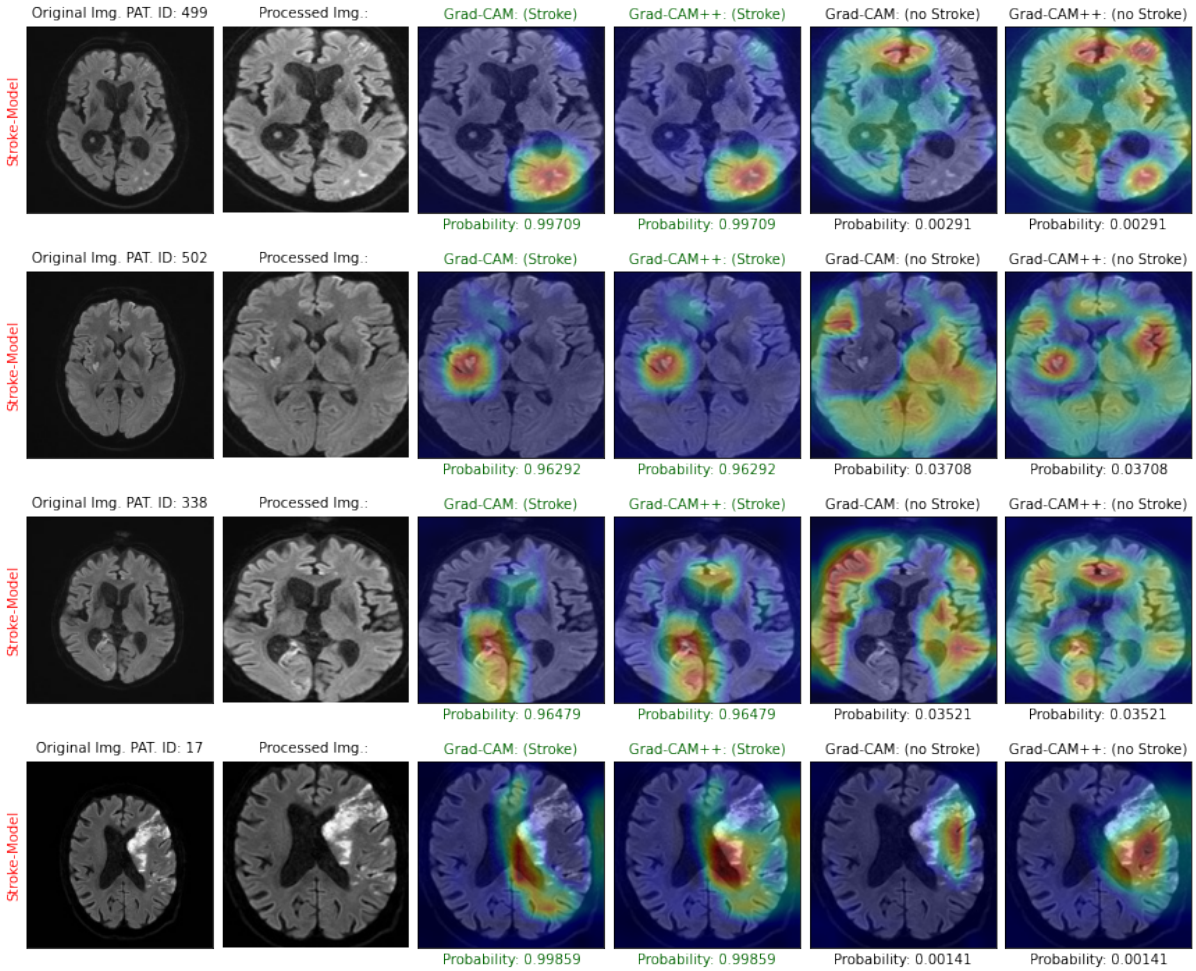

Figure 11: Linewise the graphs correspond to MRIs of different patients with a visible stroke. The first graph corresponds to the original MRI image with the patient ID annotated from left to right. The second graph corresponds to the normalized, cropped, and interpolated image. The third and fourth graphs correspond to the heatmap for Grad-CAM and Grad-CAM++ for the "stroke" class, where the x-axis label corresponds to the probability for the "stroke" class. The last two graphics correspond to the heatmap for Grad-CAM and Grad-CAM++ for the class "no Stroke", where the x-axis label corresponds to the probability for the class "no Stroke".

be made to visualize the explanations of the model. For this purpose, six graphics are plotted for each input image (cf. Figure 11). In figure 11, only images are considered in which a stroke is visible.

According to the stroke model, the first row shows an MRI of patient 499 in whom a stroke is visible with a probability of 99.7%. The original image shows bright lesions in the posterior part of the brain. Both XAI methods detect this lesion accurately. The two heat maps are similar but not identical. Other regions are only minimally highlighted. Looking at the heat maps for the class “Stroke” and “no Stroke”, it is noticeable that the heat maps for Grad-CAM do not overlap. For Grad-CAM++, the locations highlighted for the class “Stroke” are also clearly highlighted for the class “no Stroke”. This shows that the explanations of the XAI methods should always be taken with a grain of salt. For when heat maps of opposing classes overlap, the significance of class-discriminating

pixels becomes insignificant. This inconsistency is also strongly criticized in the literature (Rudin, 2019).

The examples presented in figure 11 show that the heat maps for the class "Stroke" and "no Stroke" in the Grad-CAM algorithm do not overlap at all. This observation is consistent for most of the cases considered in this paper. Also, for patients 502 and 338, it becomes clear how precisely the lesions in the brain are detected. These examples reflect the observation made for patient 499. The Grad-CAM heatmap for class "Stroke" and "no Stroke" do not overlap, while the heat map for "Stroke" detects the lesion and the heat map for "no Stroke" detects the remaining regions in the brain. In the heat maps of Grad-CAM++, there is again an overlap. Patient 17 shows a different behavior than the previous patients. The model predicts a stroke with a probability of 99%. On the MRI, a very severe stroke is visible in the brain. This is not considered class-discriminating by Grad-CAM and Grad-CAM++. The heat maps highlight the lateral ventricles where the cerebrospinal fluid is located. It is noticeable that the stroke is so severe that it has deformed this region to create an asymmetry in said zone. This may indicate that the model is learning where lesions are visible in the brain and what their effect is on other areas of the brain, such as topological deformations in the brain. Using these findings, classical interpretable statistical models can be built to test the hypotheses against new data. It should be mentioned again that these explanations should be used with caution and should always be interpreted in consultation with experts. For this particular paper, an experienced neurologist and physician was consulted.

Insights can also be gained from the misclassified images. Figure 12 shows images of patients who have suffered a stroke. The model incorrectly classifies them as images where no stroke is visible. The model is uncertain since the probability for the class "stroke" in the examples shown is still between 18% and 33%. Both Grad-CAM and Grad-CAM++ show heat maps that detect specific regions. For a non-expert, it is difficult to detect a lesion that would indicate a stroke, especially on the original images. However, consultation with an expert shows that Grad-CAM and Grad-CAM++ locate these strokes with high precision in these examples. This shows that even if the model misclassifies an input, the XAI methods can still detect the class discriminating pixels. Furthermore, it is evident from these examples that there is no overlap of the heat maps with Grad-CAM. Even with Grad-CAM++, the overlaps are very minimal.

Images taken from TIA patients all have a "no Stroke" label which by definition means they have not suffered an ischemic stroke. The model predicts a stroke on the image for some of the TIA patients. Figure 13 shows two such patients. Patient 131 and patient 135 have both been diagnosed with TIA. The model is 66% and 99% certain that a stroke is visible on the respective images. After consulting a neurologist, it is clear that the labeling for both patients is incorrect. Both patients have an ischemic stroke. The lesions affected by a stroke are recognized by Grad-CAM as well as by Grad-CAM++. Inpatient 135, the heat maps for the classes "stroke" and "no stroke" overlap in Grad-CAM++. It is unclear whether this can be interpreted since the probability for the class "no stroke" is approximately zero.

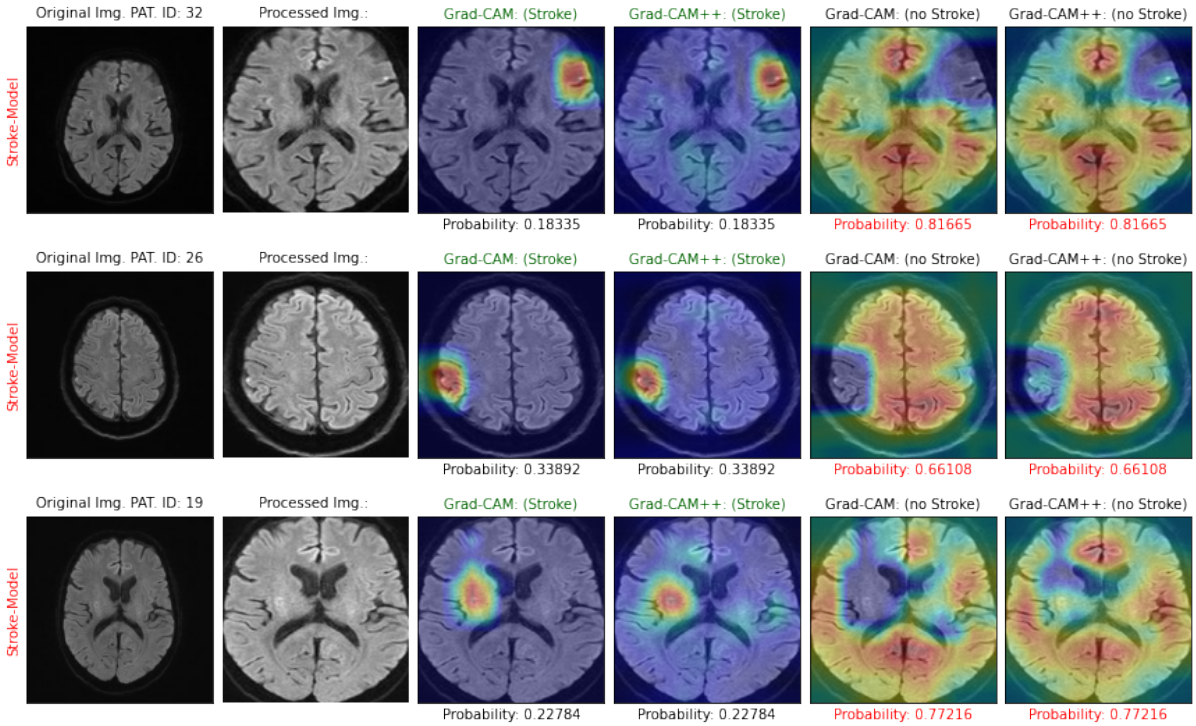

Figure 12: Images that are incorrectly classified as "no Stroke". Grad-CAM and Grad-CAM++ can detect lesions that indicate a stroke.

4.2 mRS Outcome

In a further step, it is now to be clarified whether the mRS outcome can also be predicted at the image level and whether such a model can be explained. For this purpose, the labeling method is applied as described in chapter 3.4.2. The aim is to determine whether, in the case of strokes resulting in severe disability, other regions are affected than those resulting in only mild disability.

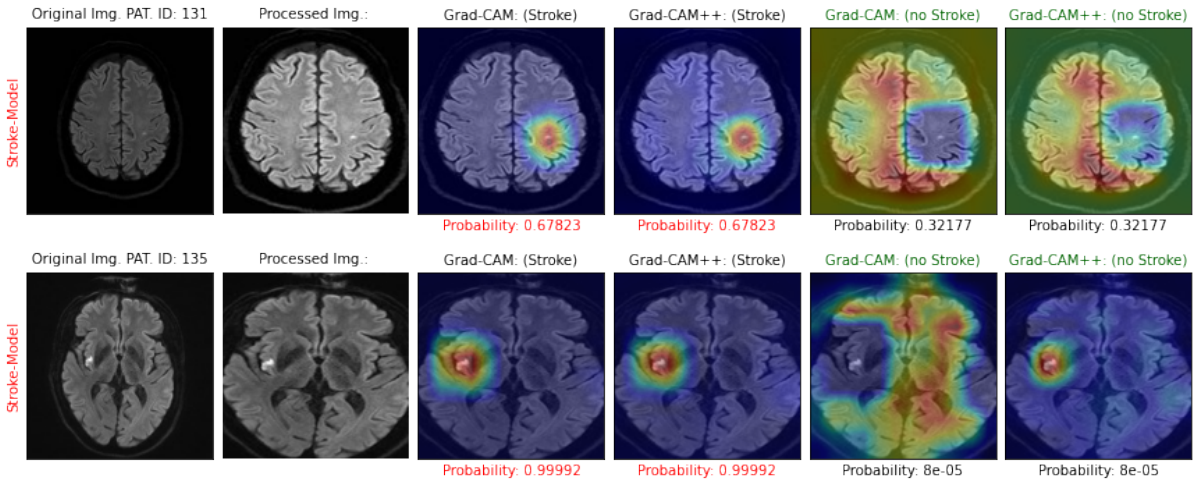

Figure 13: Patients 131 and 135 correspond to TIA patients. However, the model classifies them as stroke patients. In consultation with a neurologist, it can be determined that the model is indeed correct and that there are strokes on the MRIs.

Table 3: Performance on the test set for ordinal mRS outcome prediction.

Metric	Performance on Testset
Accuracy [95% Conf.] :	0.9206 [0.9099 0.93]
Accuracy "no Stroke" [95% Conf.] :	0.9909 [0.9859 0.9941]
Accuracy "mRS (0-2)" [95% Conf.] :	0.7018 [0.6545 0.7451]
Accuracy "mRS (3-6)" [95% Conf.] :	0.5813 [0.5125 0.647]
Negative Log-Likelihood :	0.2329

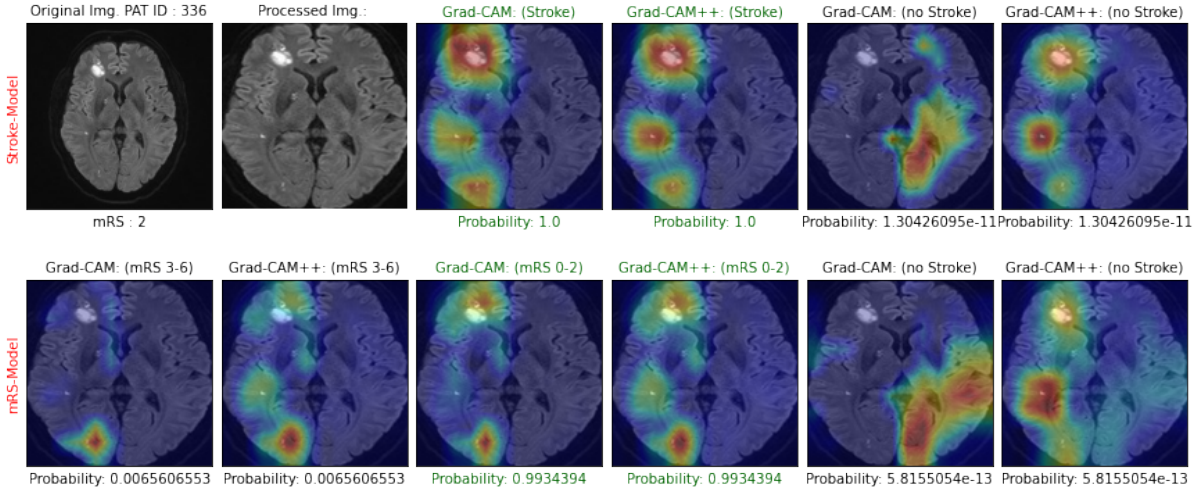

Figure 14: A stroke patient with mRS outcome two is correctly classified. The visible lesions are mainly detected as class discriminating pixels.

The classification is based on a model with the architecture described in chapter 3.4.2. The model has five convolutional blocks and global average pooling. The model is relatively simple, with less than 1.6 million parameters than the baseline model with over 16 million. The overall accuracy of the model is 92.06% [90.99% 93.00%]. As Figure 4 shows, the labels for the mRS outcome are distributed in a highly unbalanced way. This is reflected in the accuracy of the individual outcomes. Of the images that do not contain a stroke, 99.09% [98.59% 99.41%] are recognized as such. Of the images in which the patient had an mRS outcome (0-2), 70.18% [65.45% 74.51%] were recognized. The images for patients with an mRS outcome (3-6) show the worst performance. Here 58.13% [51.25% 64.70%] are recognised as such (cf. table 3).

4.2.1 Results

Figure 14 shows a variety of heat maps and key figures. The top row corresponds to the output for the stroke model. These graphs are to be interpreted as described in the previous chapters. As additional information, the patient's mRS outcome is shown in the x-axis label of the original image. The bottom row corresponds to the output for the mRS model. From left to right, the first two heatmaps correspond to the output of Grad-CAM and Grad-CAM++ for the class mRS (3-6). The center two heatmaps correspond to the heatmap for Grad-CAM and Grad-CAM++ for class mRS (0-2). The last two graphs correspond to the heat map for Grad-CAM and Grad-CAM++ class "no Stroke".

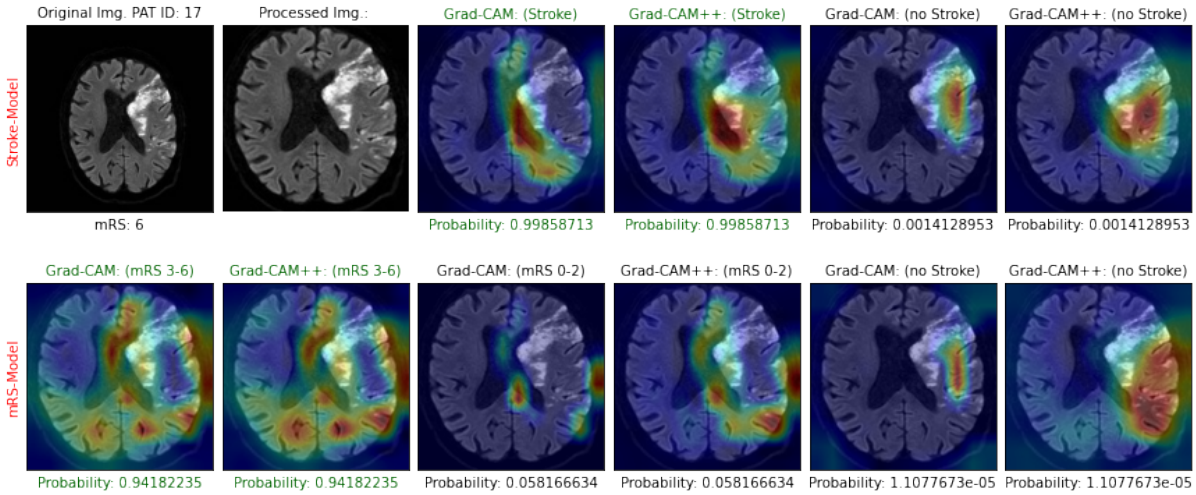

Figure 15: For severe strokes, it is not the lesions themselves but topological deformations in the brain that are considered class discriminating.

Interpreting the explanations for an ordinal outcome is difficult even for experts. The Grad-CAM and Grad-CAM++ heat maps of the mRS (3-6) and mRS (0-2) classes overlap in the posterior part of the brain. In the anterior part of the brain, the lesions are weighted more strongly for both methods. Here the model is 99% certain that the mRS outcome is 0-2. It is unclear whether the heat maps for the non-predicted classes, i.e., mRS(3-6) and "no stroke", can be interpreted at all. The probabilities for these classes are either very low or even approximately zero. Activations in the heat maps could therefore only apply to artifacts that are no longer relevant. The heatmaps for the class mRS (0-2) from both methods are not congruent to the heatmaps of the stroke model for the class "stroke", but they are similar in certain regions. It is also noticeable that the visible lesions are mainly considered class discriminating for the class mRS (0-2).

In consultation with a neurologist and physician, interesting observations were nevertheless made from which possible hypotheses can be generated. Figure 15 shows a patient with an mRS outcome (3-6). The heat maps for the stroke model of patient 17 are already explained in chapter 4.1.5. The assumption was that it was not the lesion itself that was class discriminating but the deformation of the lateral ventricle. The heat maps for the mRS model are now available as additional information. We can observe that here, in contrast to the heat maps in figure 14, not the lesion itself but different areas in the brain leading to the asymmetry of the brain are detected. We observe that the gyri (the worm-shaped outgrowths in the grey matter) swell during a stroke, causing the notches between the gyri, the so-called sulci, to disappear through compression. This closing of the sulci creates an asymmetry between the left and right hemispheres of the brain at the edges. This could be the reason why there is increased activity along the edge of the brain. Similar observations have been made in other patients. Such interpretations need to be taken with caution and no be over-interpreted but could be investigated further.

Finally, the results shown can be summarized in two different domains. From a data scientist's point of view, the heat maps for Grad-CAM and Grad-CAM++ provide insights into the neural network that would otherwise remain hidden. Whether this is a formal explanation is difficult to answer. The XAI methods shown can be used to debug or simplify models and save a considerable number of parameters, which can ultimately save

the available resources. Regarding the debugging of models, whether this does not cause two problems to be solved. On the one hand, the original problem, such as the classification model, and on the other hand, the explanation for it. It is also unclear whether the relevant features or explanations for the deep learning model have to be the same as those for experts, which leads us back to the criticism of non-falsifiability. From a neurologist's point of view, various explanations evident in the results can already be wholly understood and described as valid. In particular, in milder strokes where only selective lesions are visible on the MRI, the results show that Grad-CAM and Grad-CAM++ are very reliable and sometimes detect strokes overlooked by neurologists and physicians. The results can now be considered preliminarily since neurological hypotheses can already be generated as described earlier in this chapter, but these need further investigation.

5 Discussion and Outlook

Considering all results, it is demonstrated how XAI methods such as Grad-CAM and Grad-CAM++ help us better understand deep neural networks and at least partially free them from their black box image. With the help of Grad-CAMs, it is possible to gain insights into deep neural networks that would otherwise remain hidden. This made it possible to simplify models that were too complex by a factor of 15 and thus debug them without having to bear significant losses in performance. The XAI methods have shown that they react very differently to different architectures in terms of their explainability, whereby the dense layers, in particular, can have a suboptimal influence on the resulting heat maps. Nevertheless, the methods shown can generally be used for all CNN architectures. Using these methods, we cannot yet explain what occurs in a deep neural network, but we can determine which image regions are relevant for a specific output. Therefore, XAI methods should be considered, especially in medical image applications, to explain new models and debug existing models.

The experiments have shown that the choice of the model architecture and evaluation layer are the most critical parameters for Grad-CAM and Grad-CAM++, as others have observed (Zhang et al., 2021, Pereira et al., 2018). From the results for binary classification at the image level, it could be shown that the methods used to localize lesions in stroke patients mostly agree with the assessment of neurologists and physicians. It could also be shown that even if the model is misclassified, Grad-CAM and Grad-CAM++ can still detect lesions in the brain that indicate a stroke. The model and the Grad-CAMs were able to detect ischemic strokes in TIA patients who were misclassified. This shows that the application of artificial neural networks and XAI methods could serve as a tool for physicians and experts in the future to minimize misdiagnosis. The mRS model has shown that it can recognize different severities of disabilities from a stroke on an image. The performance of the mRS model needs to be further improved. In particular, the performance for patients and thus images with the mRS class (3-6) was severely lacking in the data, resulting in poor accuracy. The development in this area should move towards a 3-D model so that predictions are no longer generated at the image level but at a patient level, although more data may have to be collected here. The resulting heat maps of the XAI methods for the mRS model show that visible lesions are primarily detected. In some cases, the compression on the lateral ventricles generated by severe strokes has been shown to cause deformations resulting in asymmetries in the brain. Grad-CAM and

Grad-CAM++ partially detect these deformations. This may indicate that the mRS model considers other features relevant in severe strokes than in mild strokes. The explanations should always be taken with caution and not over-interpreted. It is also essential that experts are consulted to interpret the explanations, as is done in this paper. In conclusion, it can be said that the methods shown provide promising results in the area of explainable artificial intelligence when applied to medical image analysis and that it is worthwhile to conduct further research in this direction.

The central issue in this work is the two-dimensionality of the models. Since no three-dimensional CNNs were trained, the spatial information is lost. The loss of this information is especially problematic for the mRS prediction. The relatively poor data situation is also suboptimal since different mRS outcomes are strongly underrepresented in the data. To address these problems, future work on this topic should focus on 3-dimensional CNN models to retain the spatial information of the MRI. Thus, post-model XAI methods could be applied to these 3D models. First attempts at this idea can already be found in the appendix. However, more in-depth experiments could not be carried out due to time constraints. These investigations should be made, however, in work that follows this.This work aimed to apply a post-model method to a pre-trained deep learning model to explain individual predictions. This goal was achieved, although in an indirect way. The applied methods showed that the pre-trained models are too complex, and less deep models are needed to explain the predictions. With the help of experts, it was possible to interpret individual explanations of Grad-CAM and generate new preliminary neurological hypotheses.

Acknowledgements

Thanks to Prof. Dr. Susanne Wegener from the Dept. of Neurology at the University Hospital Zurich for her insightful neurological interpretations. Thanks to Prof. Dr. Beate Sick and Dr. Helmut Grabner, who supervised this work and provided me with helpful feedback.

References

• Feigin VL, Lawes CMM, Bennett DA, et. al., 2003. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. Lancet Neurol 2:43–53

• Kraft, P., Nieswandt, B., Stoll, G., et. al., 2012. Akuter ischämischer Schlaganfall. Nervenarzt 83, 435–449 . https://doi.org/10.1007/s00115-011-3368-6

• Saver, J. L., 2015. Time is Brain - Quantified. Stroke, Volume 37, pp. 263-266.

• Chilla GS, Tan CH, Xu C, Poh CL., 2015. Diffusion-weighted magnetic resonance imaging and its recent trend-a survey. Quant Imaging Med Surg. 2015 Jun;5(3):407-22. doi: 10.3978/j.issn.2223-4292.2015.03.01.

• Ker, J., Wang, L., Rao, J., & Lim, T., 2018. Deep learning applications in medical image analysis. IEEE Access, 6, 9375 9389. https://doi.org/10.1109/ACCESS.2017.2788044

• Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K. R., Dähne, S., & Kindermans, P. J., 2019. iNNvestigate neural networks! Journal of Machine Learning Research, 20.

• Herzog L, Murina E, Dürr O, Wegener S, Sick B., 2020. Integrating uncertainty in deep neural networks for MRI-based stroke analysis. Med Image Anal. 2020 Oct;65:101790. doi: 10.1016/j.media.2020.101790.

• David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller, 2010. How to explain individual classification decisions. Journal of Machine Learning Research, 11(Jun):1803–1831, 2010.

• Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Vi'egas, und Martin Wattenberg, 2017. Smoothgrad: Removing noise by adding noise. arXiv:1706.03825

• Zeiler M.D., Fergus R. 2014. Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T, 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53

• Jost T Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, 2015. Striving for simplicity: The all convolutional net. In ICLR (workshop track), 2015.

• Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, Klaus-Robert Müller, 2017. Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recognition, Volume 65, 2017, Pages 211-222, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2016.11.008.

• Pieter-Jan Kindermans, Kristof T. Schtt, Maximilian Alber, Klaus-Robert Mller, Dumitru Erhan, Been Kim, and Sven Dhne, 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. In International Conference on Learning Representations, 2018.

• Mukund Sundararajan, Ankur Taly, and Qiqi Yan, 2017. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning (ICML), pages 3319– 3328, 2017.

• Avanti Shrikumar, Peyton Greenside, Anshul Kundaje, 2019. Learning Important Features Through Propagating Activation Differences, Oct 2019.

• Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W, 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140

• Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Aug 2016.

• Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, 2015. Learning Deep Features for Discriminative Localization. arXiv:1512.04150

• Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Dec 2019.

• A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, 2018. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839-847, doi: 10.1109/WACV.2018.00097.

• Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, Garth Slaney, 2021. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging, Journal of Neuroscience Methods, Volume 353, 2021, 109098, ISSN 0165-0270, https://doi.org/10.1016/j.jneumeth.2021.109098.

• Maxim Kan, Ruslan Aliev, Anna Rudenko, Nikita Drobyshev , Nikita Petrashen, Ekaterina Kondrateva, Maxim Sharaev, Alexander Bernstein and Evgeny Burnaev, 2020. Interpretation of 3D CNNs for Brain MRI Data Classification

• Pereira S., Meier R., Alves V., Reyes M., Silva C.A., 2018. Automatic Brain Tumor Grading from MRI Data Using Convolutional Neural Networks and Quality Assessment. In: Stoyanov D. et al. (eds) Understanding and Interpreting Machine Learning in Medical Image Computing Applications. MLCN 2018, DLF 2018, IMIMIC 2018. Lecture Notes in Computer Science, vol 11038. Springer, Cham. https://doi.org/10.1007/978-3-030-02628-8_12

• Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

• Matthew L. Leavitt, 2020. Ari Morcos, Towards falsifiable interpretability research. arXiv:1909.12072

• Mihaela van der Schaar, 2020. ICML 2020: Machine Learning for Healthcare: Challenges, Methods, and Frontiers.

• Samek W., Müller KR., 2019. Towards Explainable Artificial Intelligence. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_1

• David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba, 2020. Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences.

• Lapuschkin, S., W¨aldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking clever hans predictors and assessing what machines really learn. Nature Communications 10, 1096

• Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions. In: International Conference on Machine Learning (ICML). pp. 1885–1894 (2017)

• Dr. H. L. Lutsep MD, G. W. Albers MD, A. Decrespigny Ph.D., G. N. Kamat MD, M. P. Marks MD, M. E. Moseley Ph.D., 2004. Clinical utility of diffusion-weighted magnetic resonance imaging in the assessment of ischemic stroke, https://doi.org/10.1002/ana.410410505

• van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J., 1988. Interobserver

agreement for the assessment of handicap in stroke patients. Stroke. 1988 May;19(5):604-7. doi: 10.1161/01.str.19.5.604. PMID: 3363593.

• I.S. Fernández, E. Yang, P. Calvachi, M. Amengual-Gual, J.Y. Wu, D. Krueger, H. Northrup, M.E. Bebin, M. Sahin, K.H. Yu, J.M. Peters, 2020. Deep learning in rare disease. Detection of tubers in tuberous sclerosis complex, PLoS One, 15 (2020), Article e0232376

• Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, Lawrence D Jackel, 1989. Backpropagation applied to handwritten zip code recognition Neural computation 541-551

• Goodfellow, et al., 2016. Deep Learning.

• Simonyan, K. & Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

• Chollet, F. & others, 2015. Keras. Available at: https://keras.io

• Oliver Dürr, Beate Sick, Elvis Murina, 2020. Probabilistic Deep Learning With Python, Keras, and TensorFlow Probability Chapter 4.

• OpenCV. 2015. Open Source Computer Vision Library.

Appendix

Code

All codes are available at https://github.com/avciidp/explainable_ai. The repository contains the following files:

- 2D_mrs_MODEL.ipynb: Code for the two-dimensional model for predicting the ordinal mRS outcome.
- 2D_stroke_MODEL.ipynb: Code for the two-dimensional model for predicting the binary stroke outcome.
- 2D_stroke_mrs_XAI.ipynb: Code for visualizing Grad-CAM and Grad-CAM++ for the mRS model and stroke model.
- 3D_stroke_XAI.ipynb: Code for visualizing Grad-CAM and Grad-CAM++ for the three-dimensional stroke model.
- Baseline_model_experiments.ipynb: Code for all experiments conducted with the baseline model.
- 342-0.25.hdf5: Model weights for mRS model.
- 96-0.21.hdf5: Model weights for stroke model.
- iNNvestigate_intro.ipynb: Introduction to de python library iNNvestigate (Alber et al.,2019).
- iNNvestigate_fashion.ipynb: Fashion mnist examples for iNNvestigate.
- iNNvestigate_fashion_ligt.ipynb Simplified version for fashion mnist examples.

Reproducibility

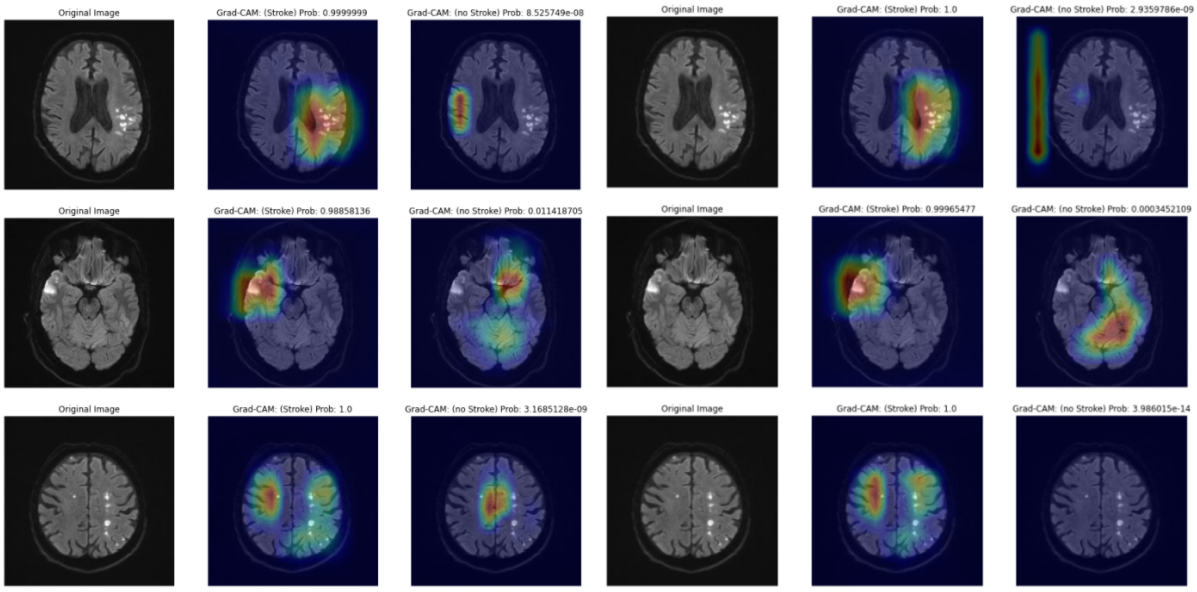

Figure 16: Reproducibility experiment: Two models are trained and validated under the same conditions. On the left, we see the Grad-CAM output for three MRIs for the first model and on the right for the second model. The Grad-CAM of the class "Stroke" is very similar for both models. For the class "no Stroke", the output cannot be interpreted as the probability of the class is approximately zero. These might represent artifacts.

Model Graph

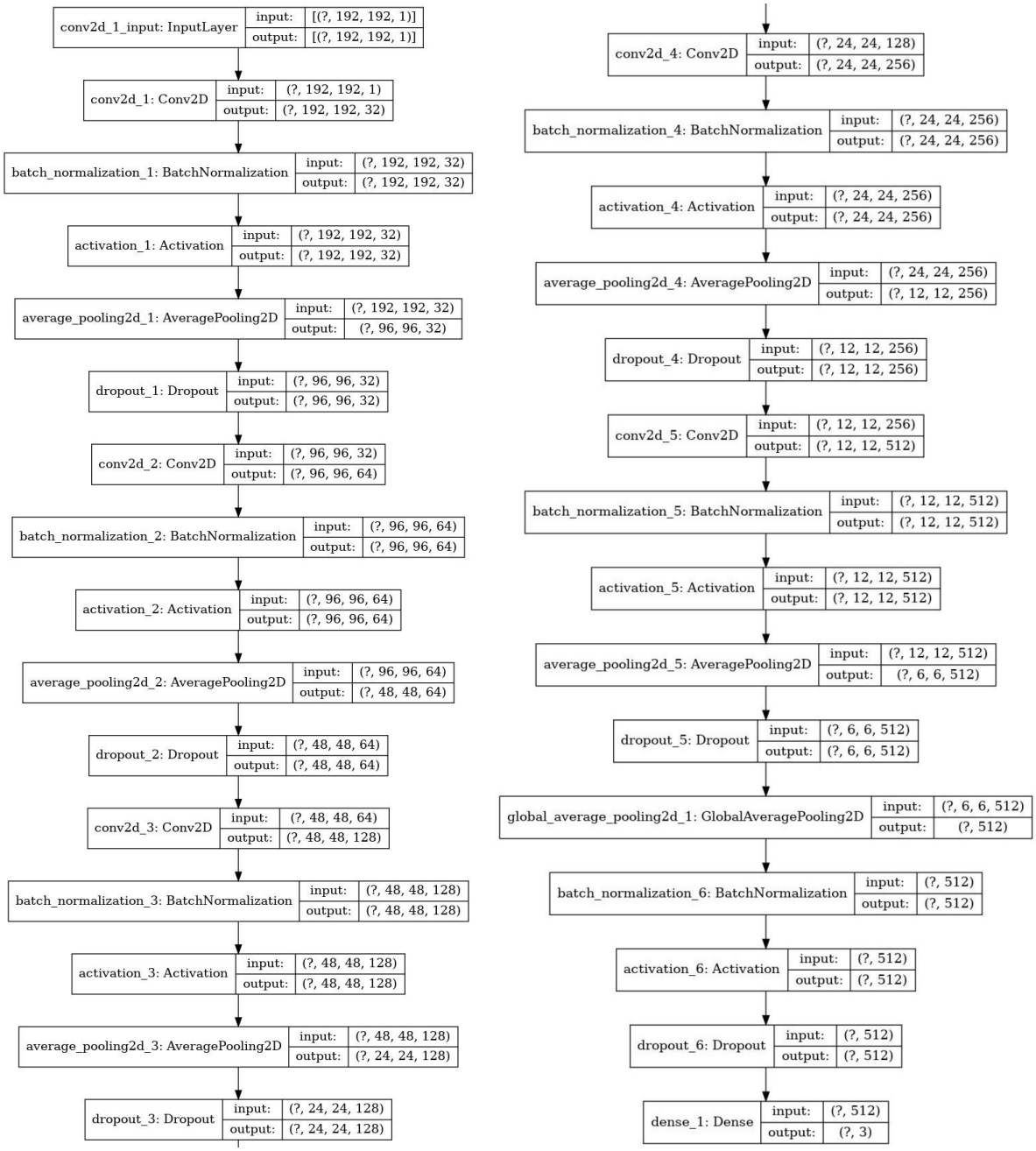

Figure 17: A visualization of the model for the ordinal mRS prediction.

3D XAI

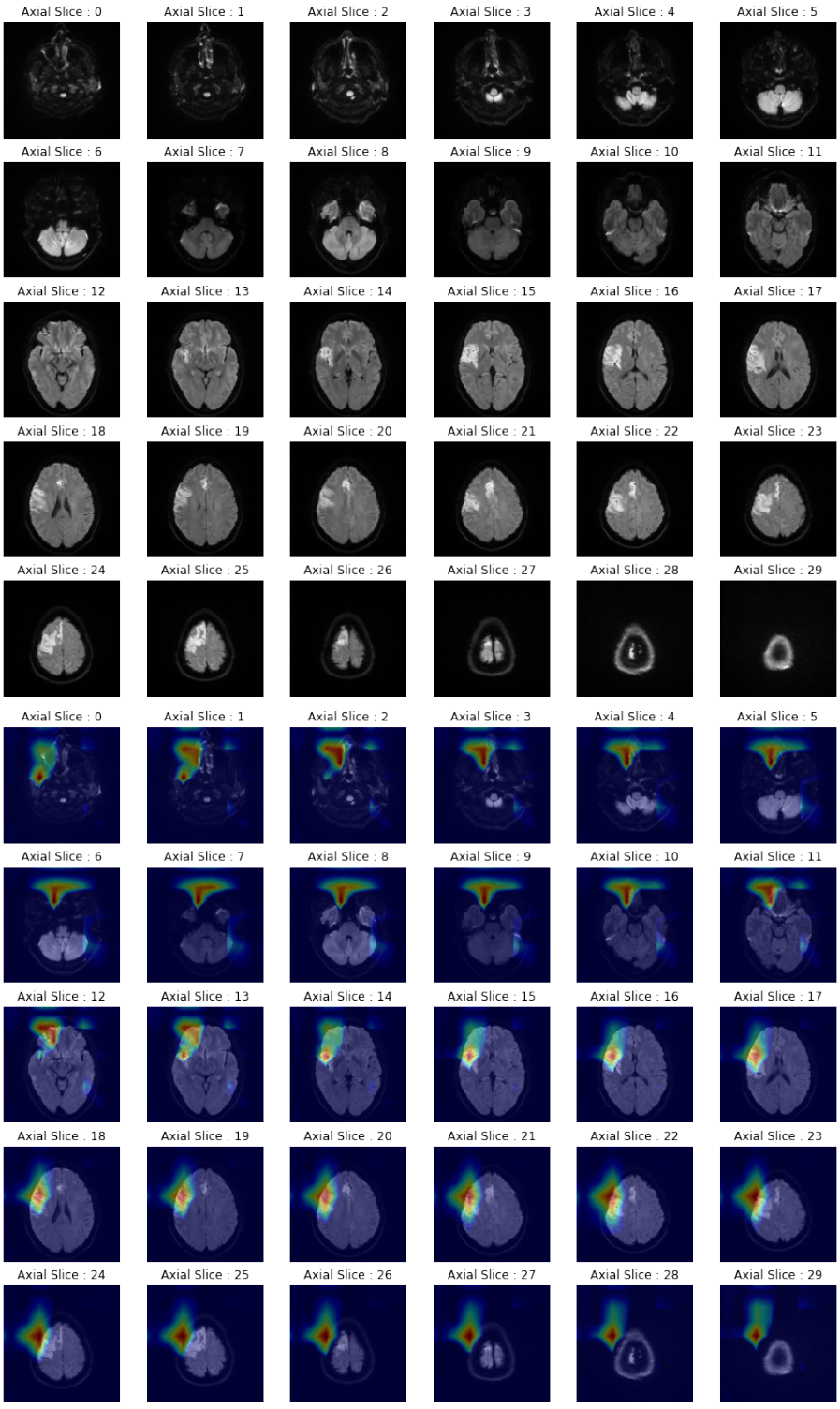

Figure 18: The top 30 images correspond to a stroke patient. The lower 30 images correspond to the output of Grad-CAM for the "Stroke" class. First experiments with 3D models show that Grad-CAM can achieve promising results in detecting lesions in the brain. The accuracy in this example varies greatly, but between Axial Slice 13 and Axial Slice 23, the detection seems to be precise.