Explainable Deep Neural Networks for MRI Based Stroke Analysis

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

Loran Avci* Beate Sick† Helmut Grabner‡

July 15, 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. Through various experiments it is shown 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 that not only can detect an existing stroke but can also 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 lesions in the brain which are caused by a stroke. The model that predicts the mRS outcome has an overall accuracy of 92%, although the data is unbalanced. It is shown that there are differences in the XAI for mRS Outcome 3-6 and mRS Outcome 0-2. The models are able to detect strokes in patients which had a transient ischemic attack (TIA) as well, and with the help of the XAI, lesions in the brain can be detected that are not recognised even by experienced neurologists. It is demonstrated that using XAI methods it is possible to simplify models without significant loss of performance. With the help of the methods presented it is possible to get a deeper look into black box models. 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 blood clot that originates from vessels in the neck, the brain, or the heart. 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 not only on clinical patient data but also on magnetic resonance imaging (MRI). MRI can not only be used for the early detection of ischemic strokes, but also to discriminate brain tumors and other useful 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). Yet artificial neural networks are still regarded as black box methods. The complex correlations that the neural network determines are usually neither explainable for domain experts nor for technical users, let alone interpretable. But it is precisely these ML algorithms that are increasingly relevant for predictive processes in critical decisions (Alber, et. al., 2019). In recent time, an increasing number of methods and approaches have been developed and introduced, which should shed light on the black boxes. These methods can be summarized under the keyword Explainable Artificial Intelligence (XAI).

In this paper, different XAI methods are evaluated. Two XAI methods which are based on class activation maps are shown in more detail. It is shown to what extent XAI methods can be used to uncover decision-relevant aspects of artificial neural networks. Several experiments with these algorithms will be conducted to show the limitations and possibilities of XAI methods. The focus of the thesis is not only on the performance of the models but also on the quality of the explanatory power of the XAI methods. MRIs and clinical patient data in tabular form from stroke subjects and a control group serve as the data basis.

1.1 Related Work

Classification of MRIs showing ischemic stroke is possible with over 95% accuracy as shown in the paper by Herzog, et. al., (2020). The authors of this paper also address 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 can be used to analyze artificial neural networks are vast. One simple method is gradient saliency maps. The size of the gradient indicates which pixels need to be changed the least in order to influence the class evaluation the most. It can be assumed that such pixels correspond to the object position in the image (David, et. al.,

2010). The SmoothGrad method averages the gradient over the number of inputs with added gaussian noise. The explanation identifies pixels that strongly influence the final decision. A starting point for this strategy is the gradient of the class score function with respect to the input image (Smilkov, et. al., 2017). DeConvNet can be seen as a method that uses the same components of a CNN as filtering and pooling but in reverse order. This method maps the labels in the direction of the pixels (Zeiler, et. al., 2014). In the guided backpropagation approach, the positive signals are backpropagated. The negative gradients are set to zero using a rectified linear unit (ReLu) function. This makes visible what the neurons in the network see (Springenberg, et. al., 2015). Deep Taylor Decomposition, as the name suggests, uses Taylor decomposition to propagate the explanations of the output back into the input layer through the network. The aim is to show which parts of the input contribute to the prediction (Montavon, et. al., 2017). PatternNet and PatternAttribution are methods that determine an estimator that determines the input signal of an output neuron. These estimators need to be trained on data, which means that the methods are not post-model methods (Kindermans, et. al., 2018). Inegrated gradients is a method in which the gradient values are cumulated along a path in the network up to the input (Sundararajan, et. al., 2017). DeepLIFT compares activations in each neuron with a reference activation and assigns a score to each neuron. This score is determined by backpropagation (Shrikumar, et. al., 2019). Layer-wise Relevance Propagation (LRP) and the numerous variations of it are a method that identifies important pixels by performing a backward pass in the neural network. Layer-wise relevance propagation applies a propagation rule that distributes the class relevance found in a particular layer to the previous layer. The layer-wise propagation rule was iteratively applied from the output back to the input, forming another possible pixel-wise decomposition. This inherits the favorable scaling properties of backpropagation (Bach, et. al., 2015). Local Interpretable Model-agnostic Explanations (LIME) is an algorithm that can explain predictions of an image classifier by highlighting the super-pixels towards a predicted class (Ribeiro, et. al., 2016). Class Activation Mapping (CAM) and its evolutions, 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).

Among countless algorithms that have been developed to analyze neural networks, there is also research that applies the behavior of neural networks to the application of medical image analysis. 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 when it comes to 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. 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, which can lead to misleading conclusions (Leavitt and Morcos, 2020).

2 Explainable Artificial Intelligence

Classical statistical models such as linear regression models have parameters which can be determined analytically or numerically. These parameters can be easily interpreted by those who use these models, which makes such models very transparent even with a large number of parameters. Deep Learning models, however, 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 a terminology for XAI for these types of models.

The trustworthiness indicates how certain a model is in a prediction. This 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 by the fact that an interpretation requires different explanations for different recipients (Van der Schaar, 2020). In the application of 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.

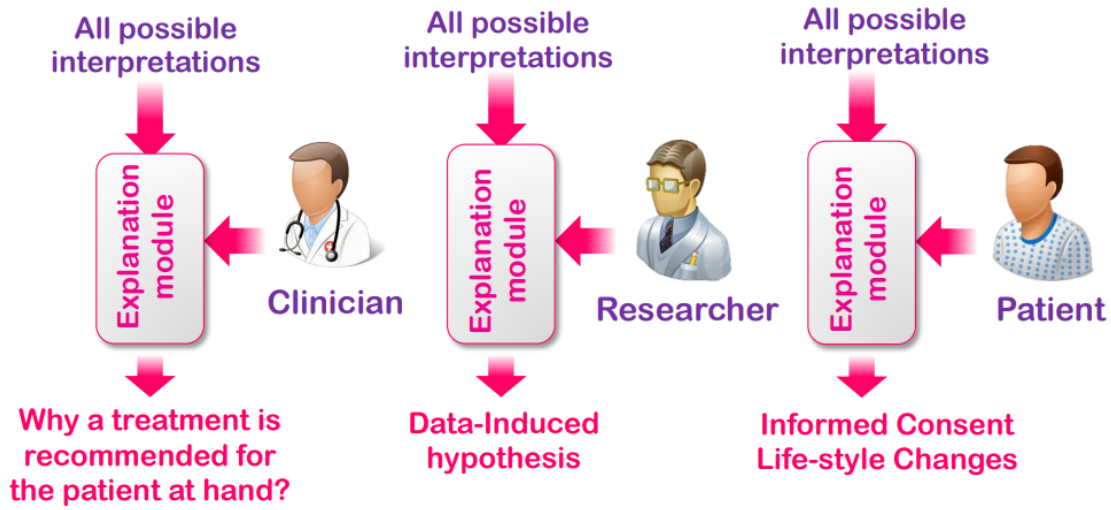

Figure 1: Explanations are dependent on the recipient. Different recipients require explanations that provide different information (Van der Schaar, 2020).

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. Research institutions, on the other hand, 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 in relation to 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 explains by highlighting the features of individual predictions that led to the classification. This could be reflected in the case of 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 biases in the model or to 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 certain 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 intention of the user of the XAI method must be clarified, i.e. what the explanation should be used for (Samek and Müller, 2019).

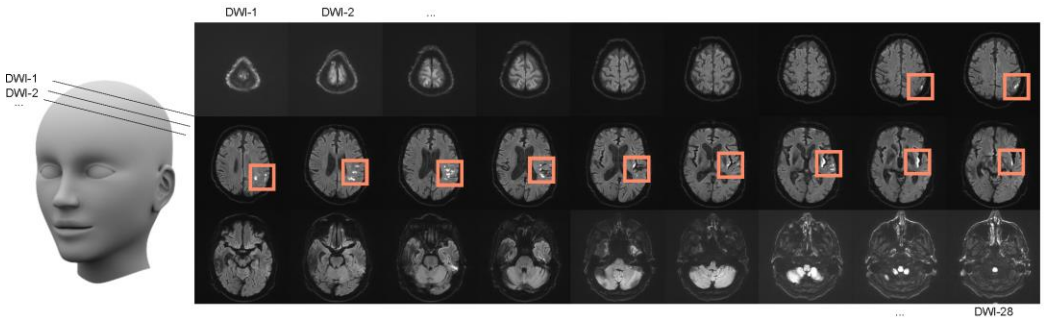

Figure 2: MRI for a patient with ischaemic stroke. The individual pictures show a two-dimensional image from an axial perspective. Strokes are usually visible on the MRI as bright anomalies (Herzog, et. al., 2020).

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 basically 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 has been shown to be superior to other MRI imaging technologies, such as T2-weighted MRI, in 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 2).

The image data is labelled by an experienced neurologist for the presence of a visible lesion to determine the ground truth. Thus, for each stroke and TIA patient, there is a “Stroke” or “no Stroke” label at 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 the 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 completely 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 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.

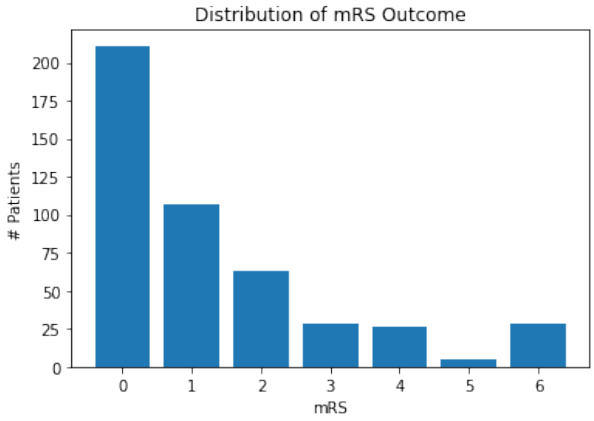

Figure 3: The distribution of mRS outcomes after 3 months is highly unbalanced.

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 as well as 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 this work, in addition to the patient ID, only the variable "mrs_3months" is included. This variable contains the information of the modified rankin scale (mRS) outcome of a patient 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 as a result of 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 3).

3.3 XAI-Methods

After evaluating different XAI methods (cf. Figure 4), it is noticeable that many of the methods often only detect edges. While this can be useful, it does not correspond to

Table 1: The modified Rankin Scale (van Swieten, et. al., 1988)

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

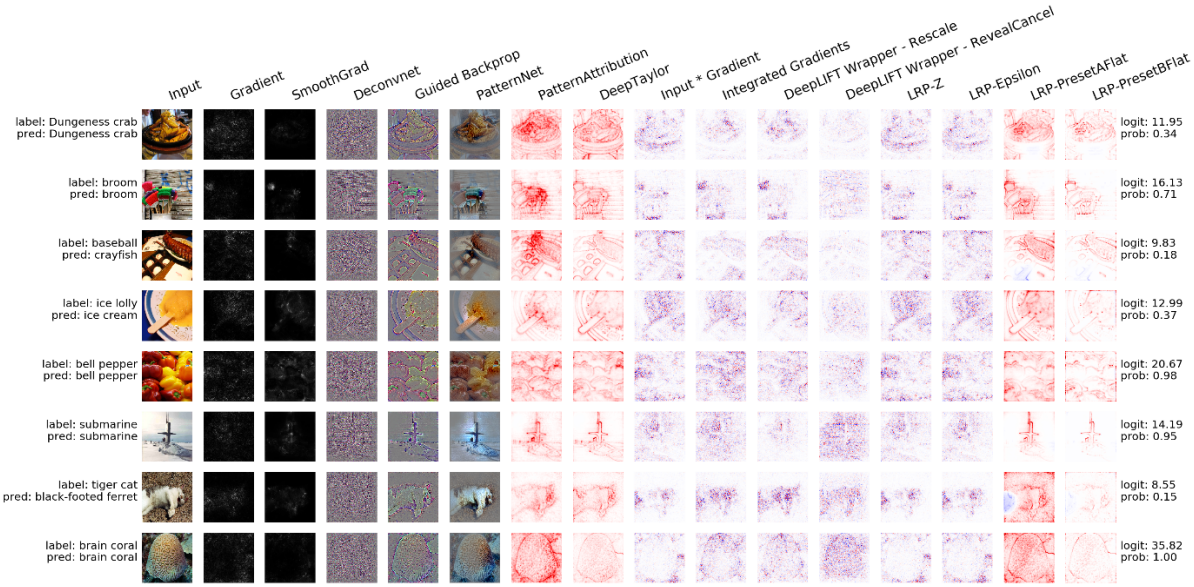

Figure 4: Overview of different XAI methods implemented in the Python library iNNvestigate (Alber, et. al., 2019).

our idea of an explanation for a neural network. CAM are an alternative approach to the methods shown in Figure 4, but this method requires a global average pooling layer (GAP) in the CNN, which limits the useage of this method (Zhou, et. al., 2016). The two further developments of CAM, namely 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). The literature research 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). A schematic visualisation of the algorithms is shown in Figure 5. The Grad-CAM and Grad-CAM++ algorithms are explained in more detail below.

3.3.1 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. Any non-one-dimensional layer can be visualised by the algorithm. 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 with respect to a feature map A^k of a covolutional 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 choosen class. Without the application of this ReLU function, the Grad-CAM would highlite regions that are not only assigned to the desired class and would therefore localise poorer. The Grad-CAM algorithm results in a heatmap that is 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).

3.3.2 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^c_k are further extended by the second and third derivatives of the gradient of the classification score:

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

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:

$$L^c_{ij} = ReLU\left(\sum_k w^c_k A^k_{ij}\right) \quad (4)$$

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

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

3.4.1 Binary Prediction on Image Level

The purpose of this model is 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 conolutional layer with a filter size of 3x3, a batchnormalisation layer, a ReLU activation, a MC dropout layer with a dropout level of 0.3 and a maxpooling layer with a pooling size of 2x2.The number of filters in the convolutional blocks increases from 32 to a maximum of 512 filters. The transition from the convolutional part to the fully connected part is done by a flattening layer. In the dense part of the model, 2 blocks are used, each consisting of a dense layer, a batchnormalization 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 crossentropy 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. 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. Based on experiments, the model architecture for binary classification will be subsequently modified several times.

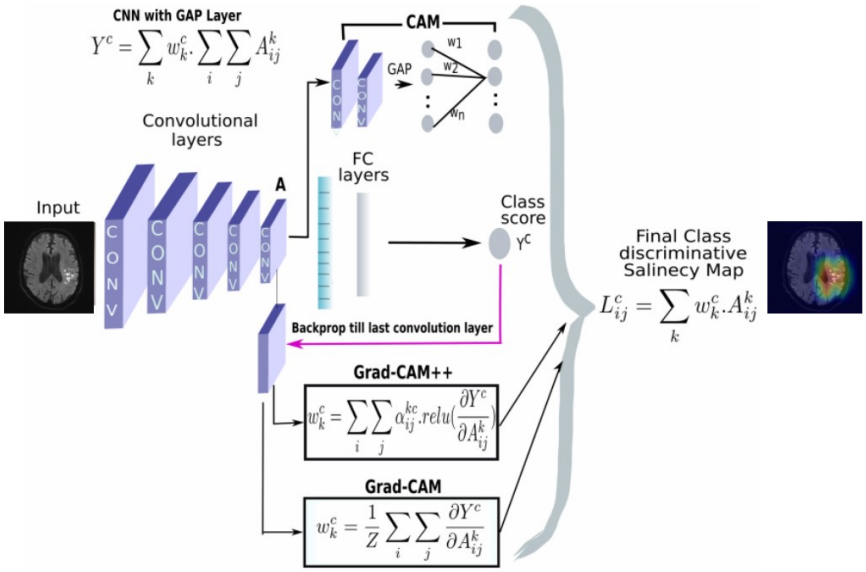

Figure 5: An overview of different CAM methods (Chattopadhyay, et. al., 2018).

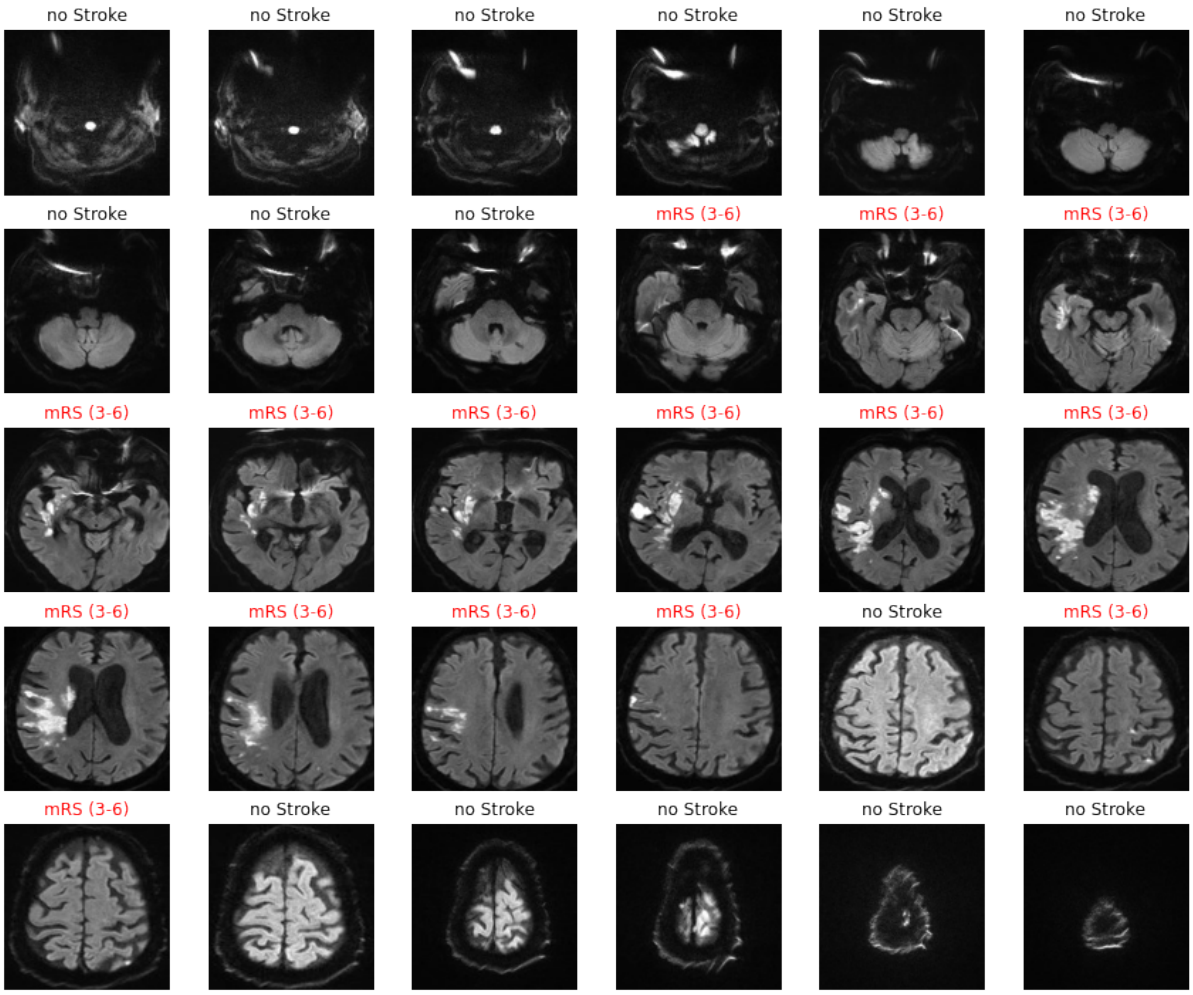

Figure 6: The figure shows how MRIs are labelled for ordinal classification. All images with a stroke label receive the corresponding binary mRS label of the patient.

3.4.2 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 range 0-6.The problem here is that these labels are only available at patient level and not at image level. Nevertheless, there is an interest in predicting the mRS outcome on the basis of a 2-dimensional image. For this purpose, a labelling method is used in this work which creates a compromise between complexity and feasibility. The ordinal mRS outcomes are first binarised. This means that 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 project the labels from the patient level to the image level, a 3-class problem is created. Depending on whether a stroke is 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, this results in the mRS outcome label 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 the label mRS (3-6). A graphical visualisation 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 5 convolutional blocks, all of which have the same structure. The blocks consist of a 2-D convolutional layer with kernel size 3x3, a batchnormalisation layer, a ReLU activation, an averagepooling 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 batchnormalisation layer, a ReLU activation and a dropout layer. This results in a model with less than 1.6 million parameters. A visualisation 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 multiclass problem the categorical crossentropy 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

The baseline model implemented by Herzog, et. al., (2020) showed that MC dropout can significantly improve the model. The test accuracy for the prediction at 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 a 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, which could indicate 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.

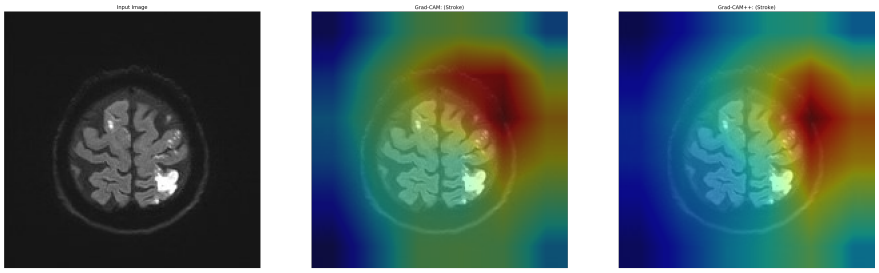

Figure 7: The input image (left) is correctly classified as a stroke by the model. Both the Grad-Cam (centre) and Grad-CAM++ (right) algorithms do not highlight the lesion.

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 leads to an improvement in terms of activations 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

A noticeable behaviour is that the Grad-CAMs often highlight similar regions that are also frequently located on the same side. Therefore, we want to investigate whether there is a misimplementation 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 a clear lesion is passed to the baseline model for prediction and then the Grad-CAM and Grad-CAM++ for the class "Stroke" are visualised. 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. If, on the other hand, 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 the 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 2). The MRIs that are close to the skullcap visualise only a small area of the brain. MRIs that image deeper layers, on the other hand, fill more pixels with relevant information. To counteract this effect, all images should be cropped so that the brain covers approximately the same surface area on all MRIs. To achieve this, a binary threshold is determined based on each individual image. This threshold is then used to algorithmically detect the contours. 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 a bicubic interpolation

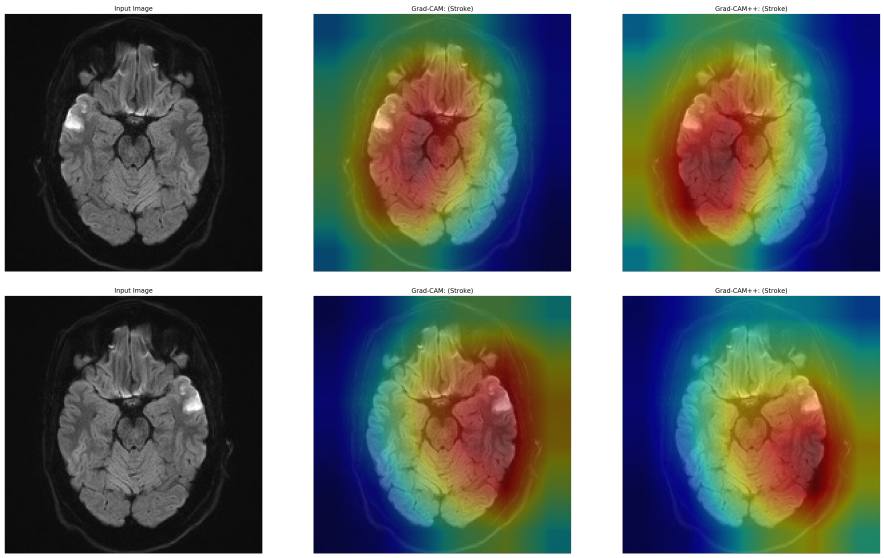

Figure 8: The experiment shows that with a mirrored image, the explanations are also approximately mirrored.

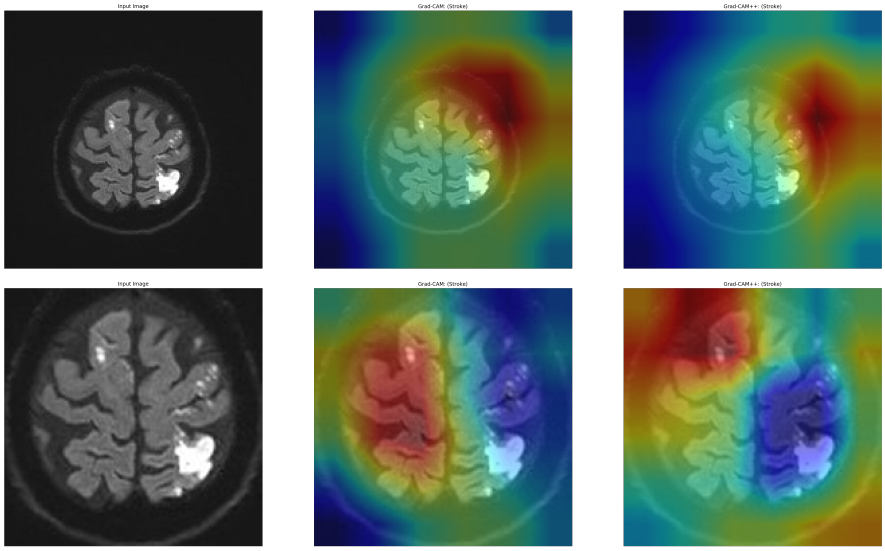

Figure 9: Cropping images leads to less activation outside of the brain in the heat maps .

(OpenCV, 2015). The results of the cropping experiments show that significantly fewer regions outside the brain are highlighted by Grad-CAM and Grad-CAM++. 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. But it is not only the regions that lie 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 the visualisation of 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 visualised in Figure 10. The resulting heatmaps show that after the first 5 convolutional layers the model considers completely different features as class discriminating than in the layers before. Especially in the last 3 convolutional layers the class discriminating pixels seem to blur. This leads us to assumption that the model is too deep and therefore too complex for this problem. The explanations interpretable to an expert are the bright spots which correspond to the lesions. These are already recognised in the first 5 convolutional layers. With this insights, the model is to be "debugged" and it is to be investigated how model performance and model 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. For all models the identical random seed is set and the same training, validation and test data is used. The influence of dense layers in the model and differences between flatten layers 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 2 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 by means of a flatten layer. While the performance is very good, the XAI visualised by the Grad-CAM is considered qualitatively poor. Therefore different architectures are trained and their performance as well as their 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 which are distributed over 5 convolutions and 2 dense layers. The transition from the convolutional part to the fully connected part is made by means of 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 which are distributed over 5 convolutions and 0 dense layers. The transition from the convolutional part to the fully connected part is made by means of a flattening 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 5 convolutional layers and 0 dense layers. The transition from the convolutional part to the fully connected part is made by means of a GAP layer. The model performs slightly worse than the baseline model. The output of the Grad-CAM is considered qualitatively

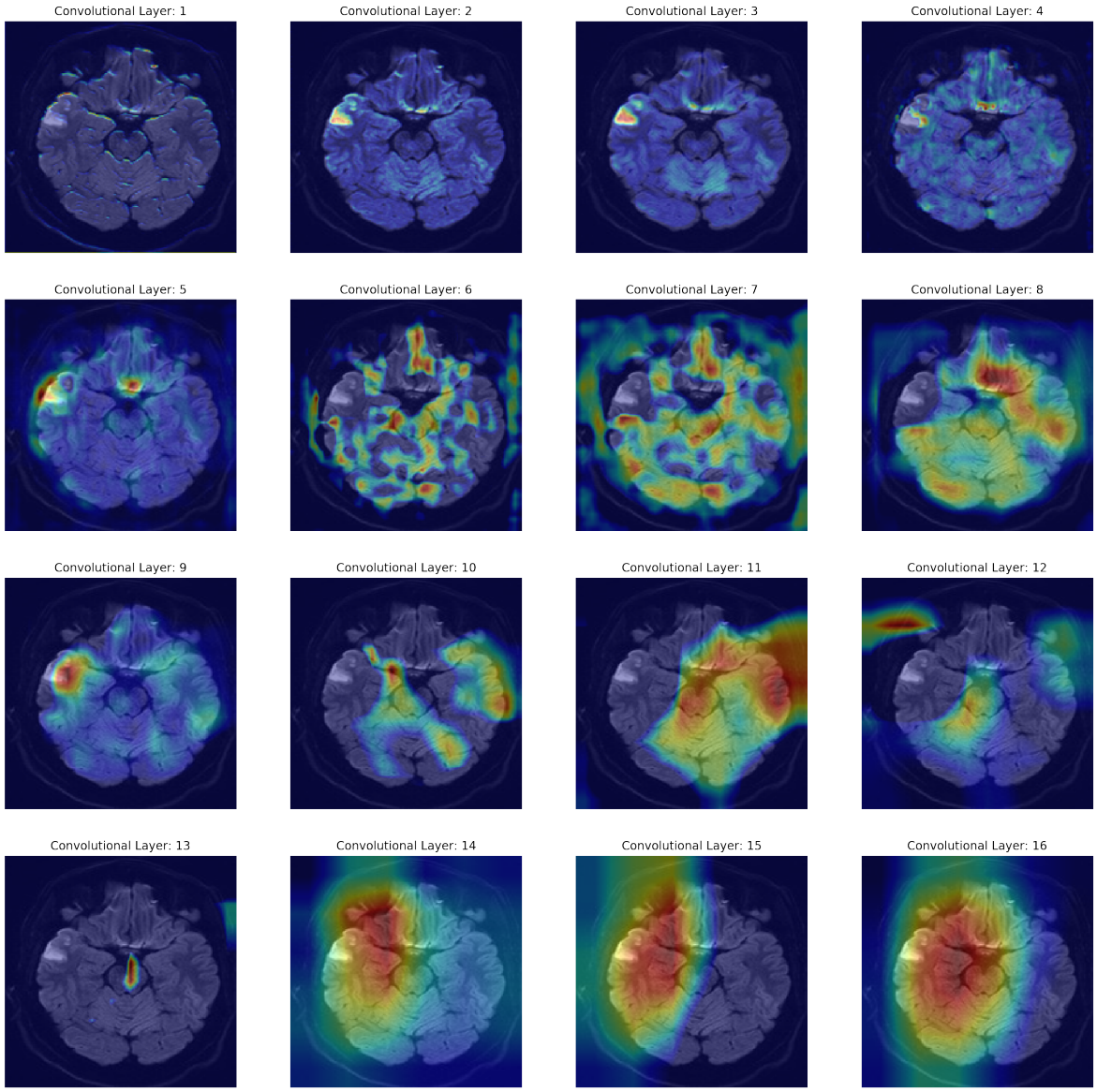

Figure 10: By iterating the evaluation layer, we observe that the class-discriminating pixels that detect the strokes are already recognised after five convolutions. These detected features become blurred with increasing complexity.

good. The "Shallow-Model" has an accuracy 89.77% [88.64% 90.79%] and has 94'000 parameters distributed over 3 convolutional layers and 0 dense layers. The transition from the convolutional part to the fully connected part is made by means of 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 specifity. Compared to the "Dense-Model", the explainability of the "GAP-Model" performs better. An evaluationlayer 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 "GAP-Model" rather favours explainability. Due to the good 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 further 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 be made to visualise 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 actually visible.

The first row shows an MRI of patient 499 in whom a stroke is visible with a probability of 99.7% according to the stroke model. 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 that are 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 criticised 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 behaviour 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 not only learning where lesions are visible in the brain, but what their effect is on other areas of the brain, such as topological deformations in the brain. Using

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

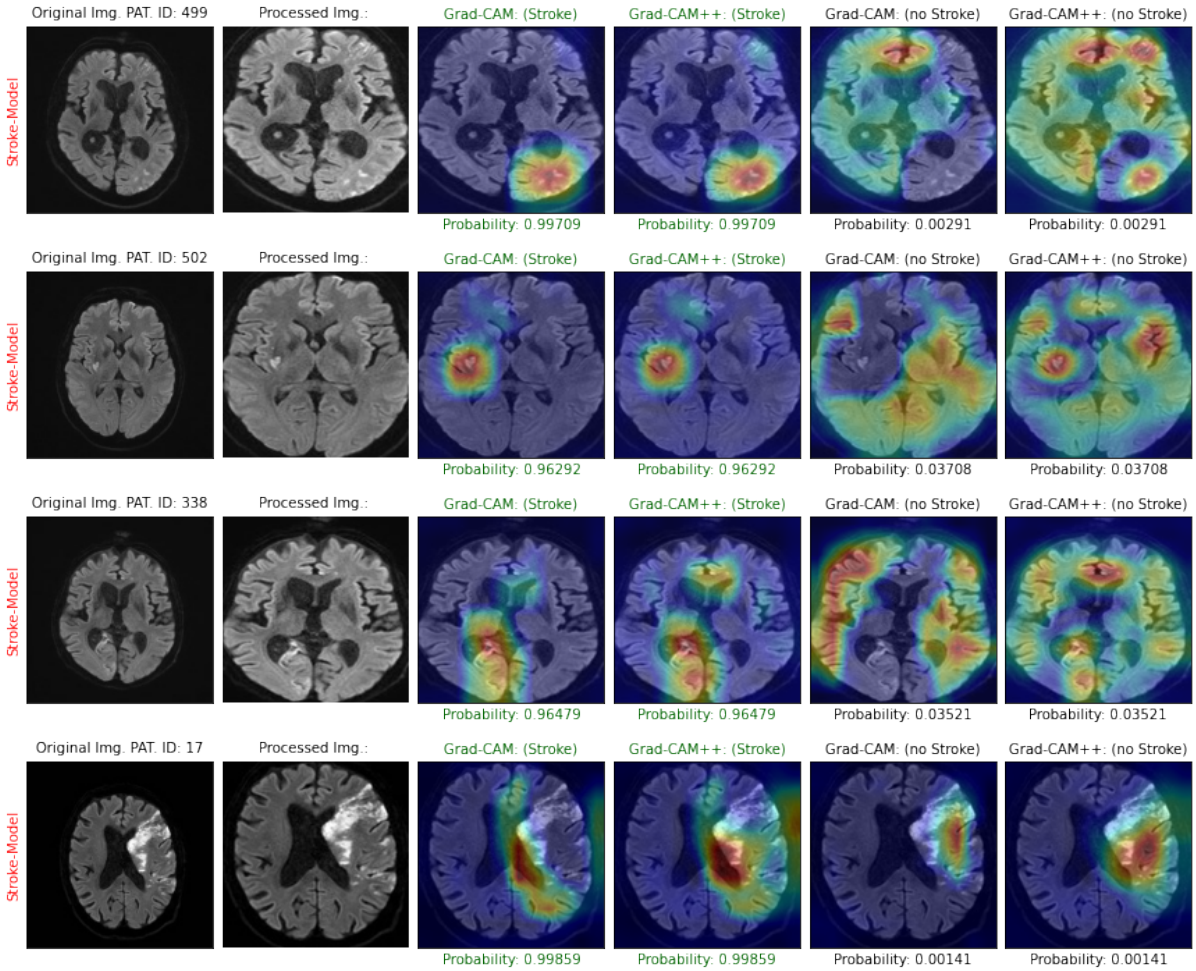

Figure 11: Linewise the graphs correspond to MRIs of different patients with a visible stroke. From left to right, the first graph corresponds to the original MRI image with the patient ID annotated. The second graph corresponds to the normalised, cropped and interpolated image. The third and fourth 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".

these findings classical interpretable statistical models can be built that could 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 particiular 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

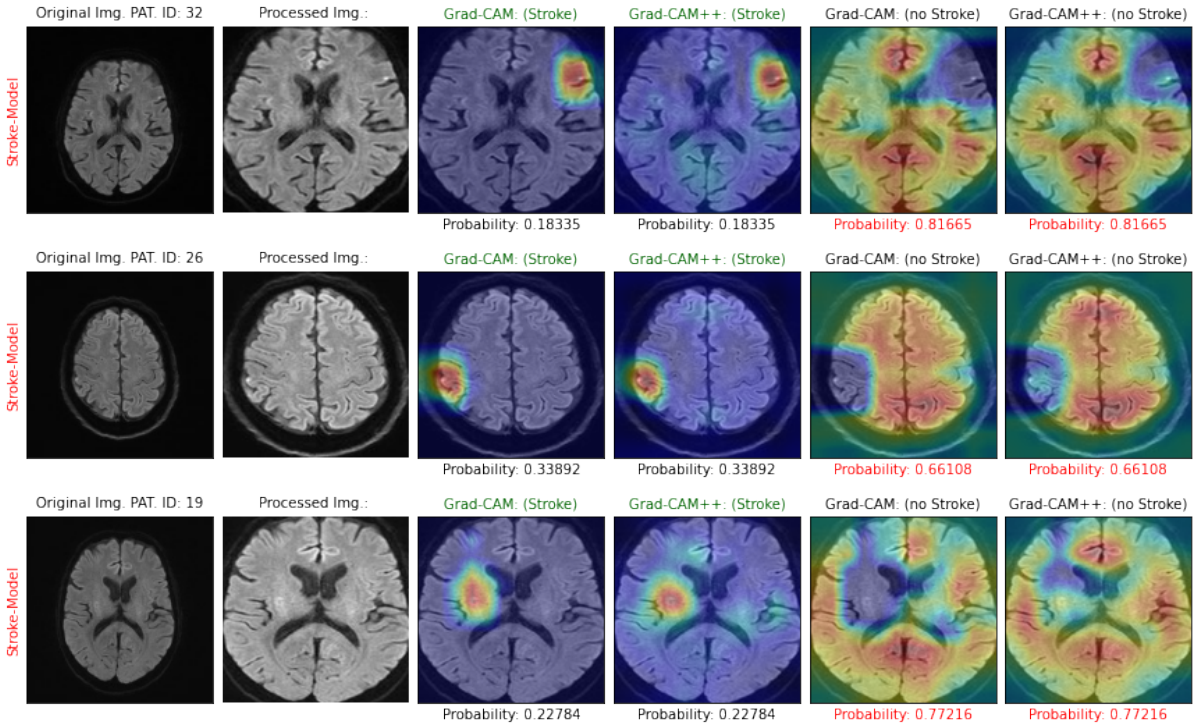

Figure 12: Images that are incorrectly classified as "no stroke". Grad-CAM and Grad-CAM++ are able to detect lesions that indicate a stroke.

misclassifies an input the XAI methods are still able to 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 as 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 labelling for both patients is incorrect. Both patients have an ischemic stroke. The lesions affected by a stroke are recognised by Grad-CAM as well as by Grad-CAM++. In patient 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.

4.2 mRS Outcome

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

The classification is based on a model with the architecture described in the chapter 3.4.2. The model has 5 convolutional blocks and global average pooling. The model is relatively

Table 3: Performance on the testset 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

simple with less than 1.6 million parameters compared to the baseline model which has over 16 million parameters. The overall accuracy of the model is 92.06% [90.99% 93.00%]. As Figure 3 shows, the labels for the mRS outcome is distributed in a highly unbalanced way. This is reflected in the accuracy for the individual outcomes. Of the images that do not contain a stroke, 99.09% [98.59% 99.41%] are recognised as such. Of the images in which the patient had an mRS outcome (0-2), 70.18% [65.45% 74.51%] were recognised. 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 2 heatmaps correspond to the output of Grad-CAM and Grad-CAM++ for the class mRS (3-6). The centre 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++ of the class "no Stroke".

Interpreting the explanations for an ordinal outcome is difficult even for experts. The

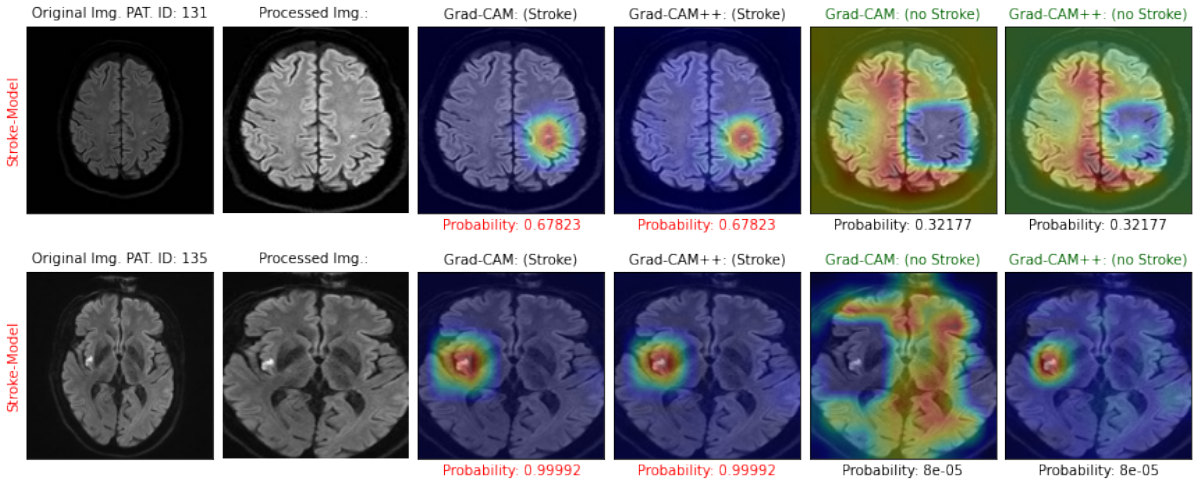

Figure 13: Patients 131 and 135 correspond to TIA patients. However, the model classifies them as stroke patients. In consultation with a neurologist, it can be determined that the model is indeed correct and that there are strokes on the MRIs.

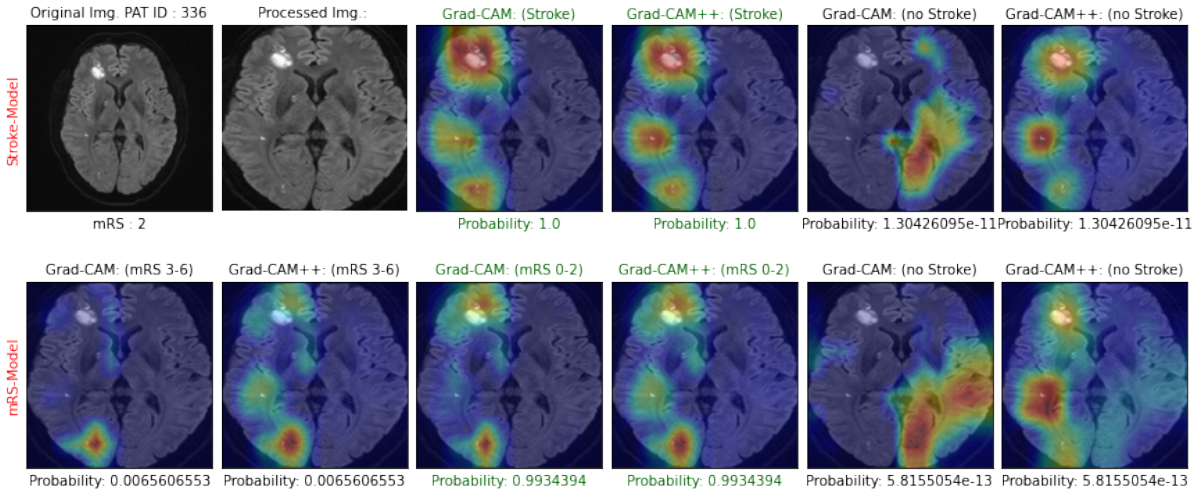

Figure 14: Stroke patient with mRS outcome 2 who was correctly classified. The visible lesions are mainly detected as class discriminating pixels.

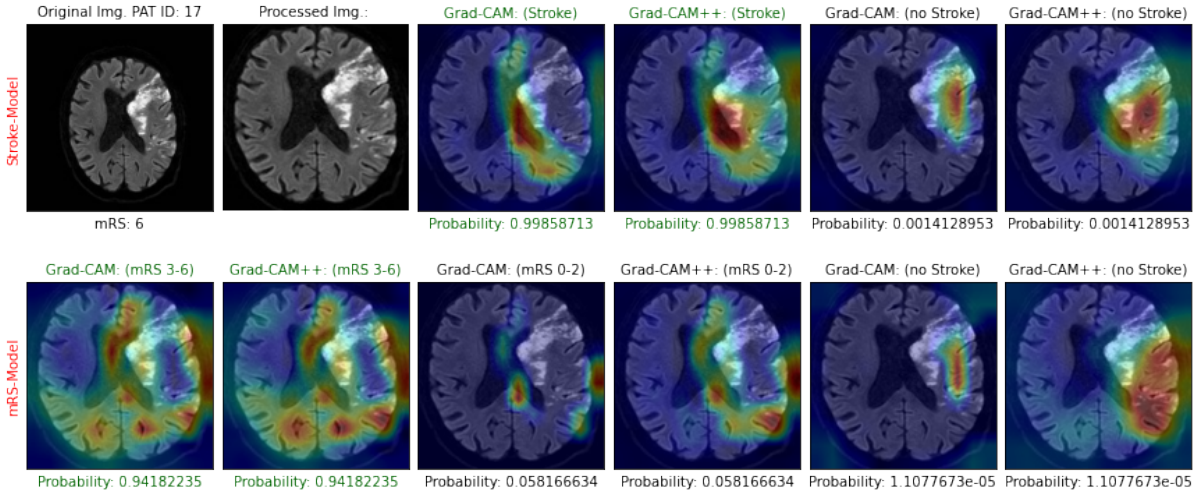

Figure 15: For severe strokes, it is not the lesions themselves but topological deformations in the brain that are considered class discriminating.

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 clearly 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 artefacts 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 for the class mRS (0-2) mainly the visible lesions are considered class discriminating.

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 mRS outcome (3-6). The heat maps for the stroke model of patient 17 are already explained in chapter 4.1.5. The assumption there 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 on figure 14, not the lesion itself but different areas in the brain leading to 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.

5 Discussion and Conclusion

Considering all results it's demonstrated how XAI methods such as Grad-CAM and Grad-CAM++ help us with a better understanding of 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 to 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. By using these methods, we cannot yet explain what exactly occurs in a deep neural network, but we can determine which image regions are relevant for a certain output. Therefore XAI methods should be considered especially in the case of medical image applications not only to explain new models but also to debug existing models.

The experiments have shown that the choice of model architecture and evaluation layer are the most important parameters for Grad-CAM and Grad-CAM++ as others have observed as well (Zhang, et. al., 2021, Pereira, et. al., 2018). From the results for binary classification at image level it could be shown that the methods used to localise 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++ are still able to detect lesions in the brain that indicate a stroke. The model as well as 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 minimise misdiagnosis. The mRS model has shown that it is capable of recognising 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 also images with the mRS class (3-6) was severely lacking in the data, which also resulted in poor accuracy. The development in this area should move towards a 3-D model so that predictions are no longer generated at image level but at 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 primarily the visible lesions are 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. These deformations are partially detected by Grad-CAM and Grad-CAM++. This may indicate that the mRS model considers other features as relvent in severe strokes than in mild strokes. Tthe 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.

Acknowledgements

Thanks to Prof. Dr. Susanne Wegener from the Dept. of Neurology at the University Hospital Zurich for her insightful neurological interpretations. Thanks also 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 PhD, G. N. Kamat MD, M. P. Marks MD, M. E. Moseley PhD, 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 two dimensional model for predicting the ordinal mRS outcome.
- 2D_stroke_MODEL.ipynb: Code for two dimensional model for predicting the binary stroke outcome.
- 2D_stroke_mrs_XAI.ipynb: Code for visualising Grad-CAM and Grad-CAM++ for the mRS model and stroke model.
- 3D_stroke_XAI.ipynb: Code for visualising 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.

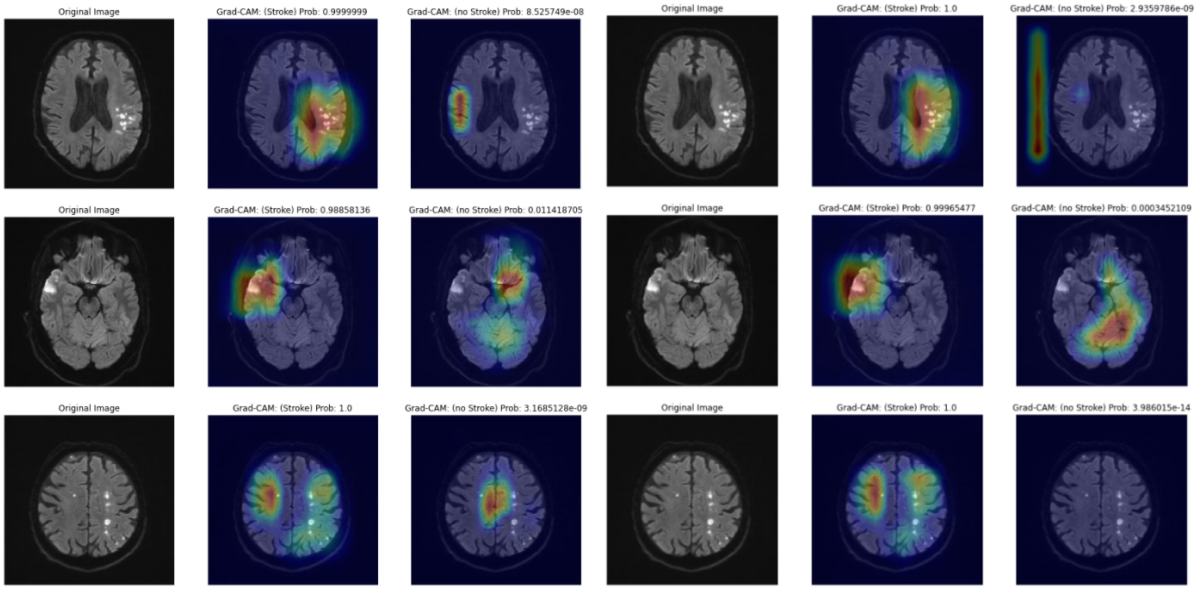

Figure 16: Reproducibility experiment: Two models are trained and validated under the same conditions. On the left we see the Grad-CAM output for three MRIs for the first model and on the right for the second model. The Grad-CAM of the class "Stroke" is very similar for both models. For the class "no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

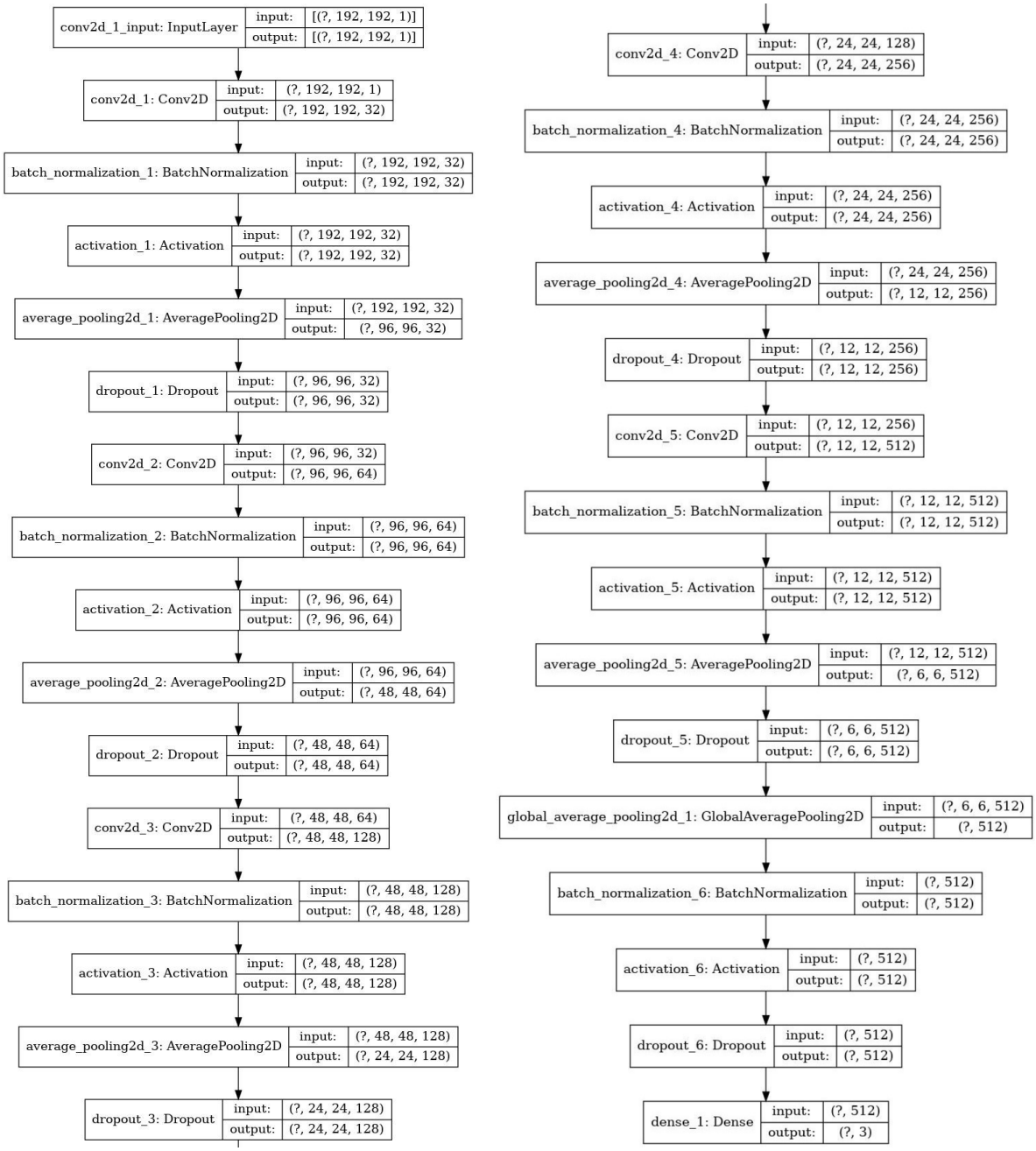

Figure 17: Visualisation of the model for the ordinal mRS prediction.

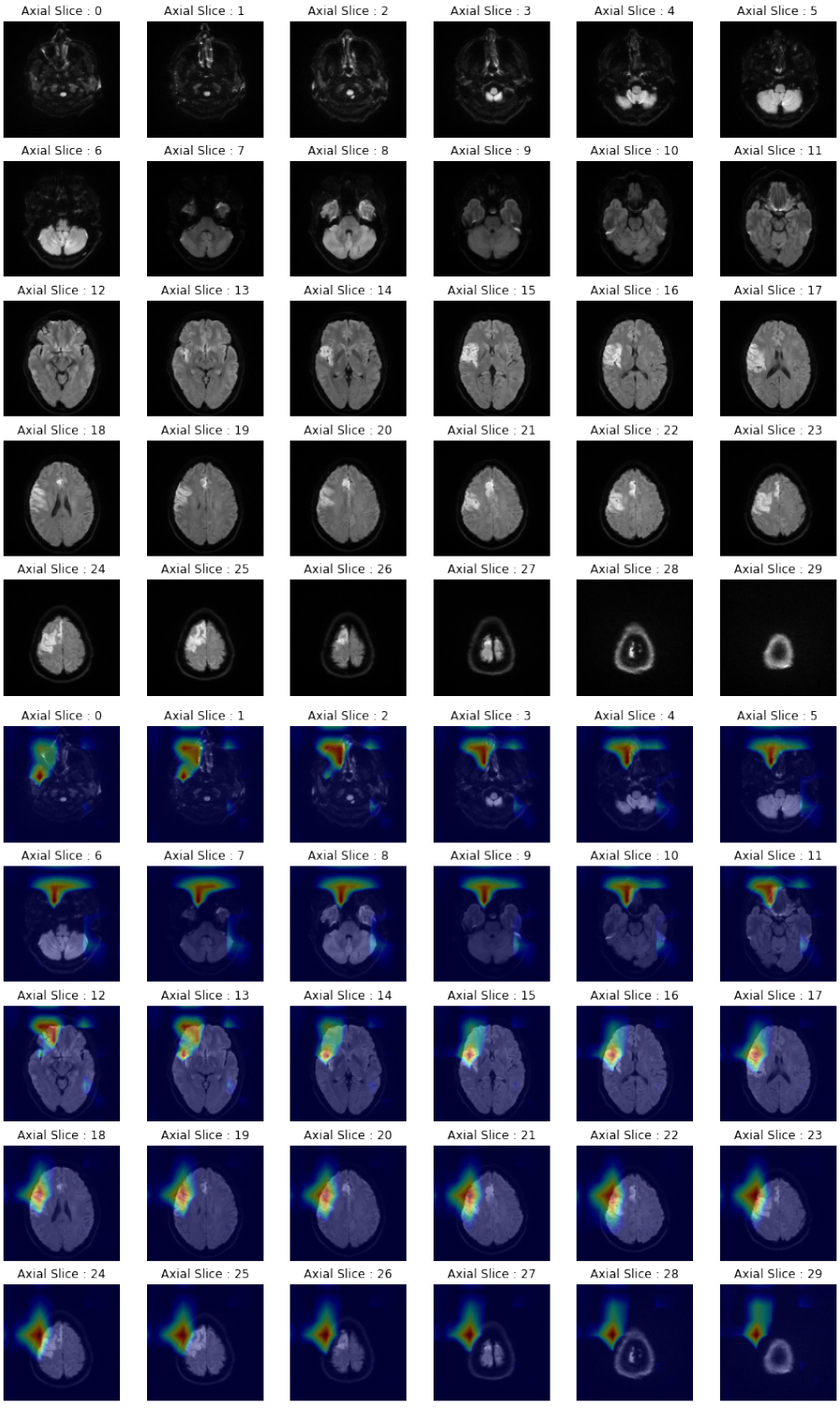

Figure 18: The top 30 images correspond to a stroke patient. The lower 30 images correspond to the output of Grad-CAM for the "Stroke" class. First experiments with 3D models show that Grad-CAM can achieve promising results in the detection of lesions in the brain. The accuracy in this example varies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