ORIGINAL RESEARCH

Improving a neural network model by explanation-guided training for glioma classification based on MRI data

František Šefčík1 · Wanda Benesova1

Received: 14 September 2022 / Accepted: 3 May 2023 / Published online: 23 May 2023 © The Author(s) 2023

Abstract  In recent years, artificial intelligence systems have come to the forefront. These systems, mostly based on deep learning, achieve excellent results in areas such as image processing, natural language processing and speech recognition. Despite the statistically high accuracy of deep learning models, their output is often based on "black box" decisions. Thus, interpretability methods (Reyes et al. in Radiol Artif Intell 2(3):e190043, 2020) have become a popular way to gain insight into the decision-making process of deep learning models (Miller in Artif Intell 267:1–38, 2019). Explanation of deep learning models is desirable in the medical domain since experts have to justify their judgments to the patients. In this work, we proposed a method for explanation-guided training that uses a layer-wise relevance propagation technique to force the model to focus only on the relevant part of the image. We experimentally verified our method on a convolutional neural network model for low-grade and high-grade glioma classification problems. Our experiments produced promising results in the way where we use interpretation techniques in the training process.

Keywords  Explainable artificial intelligence · Deep neural networks · Medical imaging · Explanation-guided training

1 Introduction

In recent years, we have been increasingly encountering the term deep learning (DL) methods. Deep neural networks (DNN) have gained importance in a wide range of applications and achieve high levels of performance in complex tasks due to their learning, reasoning, and adaptation capabilities. Despite the statistically high accuracy of deep learning methods, their output is often a "black-box" decision [13]. Therefore, methods of interpretability [21] and explainability have become important especially in life critical systems such as self-driving cars or medicine [1]. Medical experts must have an opportunity to review decisions made by models [11]. Our research focuses on the medical domain, especially on brain tumor disease. Many published works have already presented research progress in this area, e.g. the authors propose convolutional neural network (CNN) models for brain tumor detection [12, 26], features for brain tumor detection [10] and tumor segmentation using 3D U net CNN [6].

1.1 Research objective

We aim to classify low-grade glioma (LGG) and high-grade glioma (HGG) diseases from 2D Magnetic Resonance Imaging (MRI) data [15]. To achieve this objective, we first trained a CNN model. To evaluate and better understand the model’s behavior, we also implemented a method for the detection of similar cases and their explanation by visualization [18]. For this purpose, we used an inner representation of CNN with the analysis of neuron activations on hidden layers in the forward pass of the network to predict similar atlas images to an input image. On the basis of this inner representation and neuron activations on hidden layers of the CNN model, we used an explanation method based on

1 Slovak University of Technology in Bratislava, Bratislava 84216, Slovakia
the technique named BiLRP [8] to show common features between an input and sample atlas images. This explanation provides higher-order explanations that extend conventional visualization techniques. Using this method, we evaluated that our first model does not work correctly with sufficiently high accuracy and explanation by visualization showed that this model for brain tumor classification took into account also irrelevant features. We have concluded, that the model should pay much more attention to the tumor region which led us to propose the presented explanation-guided training method.

### 1.2 Our contribution

Our proposed explanation-guided training method improves the model’s prediction and forced the model to focus more on the tumor region in 2D MRI brain slices. The main steps of our work are described as follows:

- We implemented a method for explanation of the model by predicting similar cases to input image and explaining it by visualizing similar features between these images.
- We used our explanation method to derive an explanation of the trained model to reveal the pitfalls of the predictions.
- We propose a method of explanation-guided training to prioritize relevant features.
- We performed experiments to show the results of the proposed method on the trained model for glioma disease classification.

Hence, the main contribution of this article can be summarized as follows:

- We introduce a novel method of explanation-guided training of DNN to prioritize relevant features in the input during the training phase.
- While the authors in the related work [28] presented a method of explanation-guided training to dynamically finds and emphasize the features which are important, our method aims to allow domain experts to prioritize important features in the input image by segmentation mask. Hence, the proposed method is useful when we want to prefer some well-known input features, but we do not want to restrict the model only to these features.

### 2 Related work

Explainability is one of the main obstacles AI currently faces in practical implementation. The needs for XAI have been commonly described and broken down by authors in [1, 9, 23] into four main motivations: reasoning about decisions for people who will be affected by the decisions of these systems; control over the system to prevent things from going wrong; improving models; and discovery to uncover patterns that humans cannot capture. To achieve these goals, many XAI methods have emerged with different approaches for a spectrum of problems. XAI methods can be divided into two main groups: transparent models and post-hoc explainability. Transparent models are models interpretable by their design. The purpose of the method is to design transparent models that are self-explanatory. Post-hoc explainability is explaining of the "black-box" model. The vast group of post-hoc techniques has been divided into several subgroups in [3] as: explanation by simplification; feature relevance explanation; explanation by visualizing hidden abstractions; and architecture modification.

#### 2.1 Perturbation based methods and occlusion sensitivity

One of the most straightforward methods interpreting the DNN prediction is Perturbation based methods [22]. This broad category of techniques perturbs intensities in the input image and observes the changes in prediction probabilities. The main idea is that the pixels that contribute maximally to the prediction will reduce the prediction probability if they change. Occlusion sensitivity was presented in the paper Zeiler et al [31]. In this approach, they occluded a part of the image with a gray window, and they were sliding this window across the whole image. For each window position, prediction probability is monitored and captured by color range. Finally, we can see heatmap with the areas that have the greatest impact on a particular prediction. This idea was developed also in another approach named Randomized Input Sampling for Explanation (RICE) [20]. They obtain the importance map by probing the model with randomly masked versions of the input image instead of a sliding window over the input image. Local Interpretable Model-agnostic Explanations (LIME) [22] is another type of Perturbation methods that builds surrogate models around a black-box prediction to explain it. LIME trains the local surrogate model to approximate a prediction of the complex model instead of training the global surrogate model. Perturbation-based methods work well for explaining decisions but suffer from computational cost and instability to surprising artifacts.

#### 2.2 Gradient-based methods

Methods based on gradient offer another way to interpret models. These methods involve a single forward and backward pass through the network, in addition to multiple forward passes in Perturbation-based methods. Also, gradient-based methods are computationally more efficient and stable.
in artifacts compared to perturbation-based methods. One of the methods from this category is Sensitivity analysis [32] and it assumes that the most relevant input features are those to which the output is most sensitive. It attempts to explain the prediction based on the local evaluation of the gradient of the model. Sensitivity analysis mathematically quantifies the relevance of input variables \( R_i = \left( \frac{\partial F(x)}{\partial x_i} \right)^2 \). This technique is easy to implement for neural networks because the gradient can be computed using backpropagation. The purpose of applying sensitivity analysis is usually not to explain the relationships found. Instead, it is generally used to test the stability and credibility of models, as well as a tool to remove some unnecessary input attributes or as a starting point for more complex explanatory techniques such as decomposition [23]. There is some further modification of gradient’s methods such as Integrated Gradient [29], Gradient*Input [25], Guided Backprop [27] or Grad-CAM [24]. These algorithms differ in the way that gradients are modified in a backward pass. However, these techniques must deal with the problems of propagating gradients through non-linear layers.

2.3 Graph structure based methods

Another approach to explain the prediction of Deep neural networks is the explicit use of their graph structure. These methods use the following procedure: starting at the network output, the prediction is mapped in the reverse direction of the graph, so we always map the prediction to lower layers until we reach the input of the network [18].

2.4 Layer-wise relevance propagation (LRP)

Layer-wise relevance propagation (LRP) [4] is a method, that operates by propagating the prediction backward in the neural network, using a set of purposely designed propagation rules. In the paper from Papernot et al. [19], an approach that is completely different from previous methods is proposed. They apply K-Nearest Neighbour (KNN) algorithm to the representation of data learned by layers in CNN to understand model failures. The test input is compared with neighboring training points by distance in representation learned in hidden layers of CNN. Then, the confidence of the model is taken as the homogeneity among neighbour labels. We can detect adversarial examples or protect model from examples that are outside of model understanding. BiLRP [8] is a method based on second-order explanation. This method was introduced to explain the similarity between two inputs to the model. The method was inspired by LRP to bring robustness for explaining of dot product similarities. BiLRP performs a second-order deep Taylor decomposition [17] of the similarity score, which enable to examine common features that contribute to similarity on any hidden layer of DNN. Explanation techniques are usually used to explain decisions of a trained model. Authors of the recent work [28] presented explanation-guided training for task of cross-domain few-shot classification. They developed a model-agnostic explanation-guided training strategy based on the LRP method that dynamically finds and emphasizes the features which are important for the predictions. Their work shows application of explanation for training phase which effectively improves the model generalization.

3 Proposed method

3.1 Problem description

Before presenting our method for explaining and improving model prediction by interpretation technique, we introduce a task in the domain of LGG vs. HGG diagnosis classification which is a combination of tumor region segmentation and patient survival prediction. In general, LGG cells do not attack normal neighboring cells, while HGG cells attack their adjacent cells [30]. Accurate classification of gliomas is therefore an important requirement because the type of glioma has an impact on the patient’s overall survival. This problem has also been discussed in other papers [2, 7, 14].

3.2 Dataset

To train our model, we selected BraTS dataset. Dataset consists of multi-modal MRI scans of glioblastoma (GBM/HGG) and lower-grade glioma (LGG) with a pathologically confirmed diagnosis.[15] Segmented annotations are available for each volume.

3.3 Model and training

In the preprocessing phase, we used three MRI sequences (T1-weighted and T2-weighted, and FLAIR) to generate three-channel 2D tumor slices from MRI brain volumes. For each class, different numbers of slices with the largest area of tumor region were extracted to balance the dataset. In the next step, we trained the model of CNN with prepared 2D tumor slices for the classification of LGG and HGG tumors. The architecture of the network is built from four 2D convolution layers with ReLu activation function followed by Max pooling layers, ending up with two fully connected layers. Detail description of the network’s layers is shown in Table 1 (Fig. 1).

1 https://www.med.upenn.edu/cbica/brats2020/data.html.
To avoid fast overfitting, we applied a random rigid augmentation with rotation, horizontal and vertical flip transformations. Neuron dropout was also applied after the first and last max-pooling layer. Adam’s optimizer with Sparse Categorical Crossentropy loss function was used as the optimization function.

3.4 Detection of similar cases and their explanation by visualization

To explain our trained model, we implemented a method for detection of similar cases and their explanation by visualization showed in Fig. 2. This method is not an essential part of the paper, but forms the basis for the proposed method in the following Sect. 3.6. The method is based on prediction similar atlas images based on their features in hidden layers, inspired by Deep k-Nearest Neighbors [19]. The method uses KNN classifier to find the most similar images to the input from neurons activations on every hidden layer. The output of the method is used to allow domain experts to investigate pairs of similar cases in network predictions on every layer.

Here, domain experts can ask the question: What led to the prediction? The answer is an explanation technique to clarify what the network sees in the related images. Experts should be able to address how the images are related to each other. We call this type of explanation Contrastive explanations [16]. Here we proposed to use higher-order explanations to show domain experts what features between an input image and an image from ground truth cases are commonly similar. To bring these higher-order explanations to MRI data classification, we introduce using of BiLBP [8] technique. Similar atlas images to the input image are visualized in our method. Therefore, we can represent the classification of hidden layer features using KNN as a similarity model, and the BiLBP method is suitable for this type of problem. We applied BiLBP to the features of all selected hidden layers. We consider the selection of features of the hidden layer as crucial for the explanation and their subsequent interpretation for the user.

3.5 Visualization by BiLBP explanation

Figure 3 shows the result of our method for similar cases detection and their visualization by BiLBP explanation. An important conclusion follows from these explanations. The model should point to areas of brain tumors; instead, the model focuses on areas that are not primarily related to brain tumors. This suggests that the model did not learn the correct information which we would consider the most relevant to the model, despite the model’s reasonably good accuracy. These observations led us to develop a method to address this problem.
3.6 Novel method of explanation-guided training

The proposed method introduces a novel way of training the neural network. It brings the type of explanation-guided training where we can use complementary information in form of a segmentation mask to force the model to focus only on the relevant part of the image. The method does not require changes in the architecture of the model, but it modifies the loss function in the training phase of the model. One training phase involves two following steps illustrated in Fig. 4 as blue and red paths. For each training iteration: 1st step (blue) The input passes forward through the network to obtain prediction value \( \text{pred} \) as the output. 2nd step (red) involves interpretation method LRP to obtain relevance score \( R \) from output \( \text{pred} \). LRP is initiated from a neuron corresponding to the true label \( y \). The \( R \) score tells how much each pixel of the input image contributes to the prediction. With the segmentation mask of the relevant region from the input image, we can observe if relevant pixels with high score \( R \) are present in the region. Thus, we can quantify how much attention the network gives to the relevant region. We used this information to modify the training phase and to design a new loss function.

Algorithm 1 Updated algorithm of model training

```
repeat
  batch ← generateBatch()
  for data ∈ batch do
    (image, y, mask_tumor, mask_brain) ∈ data
    \( \text{pred} \) ← model(image)
    \( R \) ← LRP(model, \text{pred})
    \( R_{mask} \) ← \( \sum \text{takePixels}(R, \text{mask_tumor}) \)
    \( R_{brain} \) ← \( \sum \text{takePixels}(R, \text{mask_brain}) \)
    tumorLRPScore ← \( \frac{R_{mask}}{R_{mask} + R_{brain}} \)
    \( \text{LRPLoss} \) ← \( \frac{\text{CategoricalCrossEntropy} \left( \text{pred}, y \right)}{\text{tumorLRPScore}} \)
  end for
until enough train iteration
```

We applied the proposed method to the aforementioned task of tumor disease classification. In addition to evaluating how accurately the model classifies between the LGG and HGG classes, the newly proposed loss function also uses an LRP interpretation technique to penalize the model if it does not target the tumor region during prediction. The loss function is named \( \text{LRPLoss} \) and can be written as:

\[
\text{LRPLoss} = \frac{\text{CategoricalCrossEntropy}(\text{pred}, y)}{\text{TumorLRPScore}(\text{pred}, x_{seg})}
\]

where \( \text{CategoricalCrossEntropy} \) is calculated from \( \text{pred} \) score as output of network and \( y \) as true label. This loss value is penalized with \( \text{TumorLRPScore} \) which describes how much attention is paid to the tumor region. When the score is high then the most relevant pixels are situated in the relevant region of tumor. The score is calculated as:

\[
\text{TumorLRPScore} = \frac{R_{mask}}{R_{mask} + R_{brain}}
\]

where \( R_{mask} \) is the sum of LRP relevances inside of the tumor mask while \( R_{brain} \) is the sum of LRP relevances of the whole brain mask without the tumor region. Different types of LRP methods can be used. In some of them, the relevances can acquire positive and negative values, then we count only positive values. We also use a version of \( \text{TumorLRPScore} \) where both its components are normalized by the size of their regions. After the observation, we learned that the first version of the scores describes the property in a better manner. Subsequently, we were training the models with the designed loss function.

In conclusion, we have designed a novel loss function, which is used to modify the training phase. In ordinary training, we use images with their labels. In the proposed training, we need images, segmentation masks of tumors, and labels.

4 Experiments and results

To evaluate our method, we proposed two experiments. In the first experiment, we applied the proposed method to our trained model where we compare different levels of penalization in \( \text{LRPLoss} \). In the second experiment, we applied our method to the training of state-of-the-art solution from Subhas et al. [5] Their solution is comparable with our model.
with similar data pipeline where slices are extracted from volume as 2D images. They achieved an Accuracy 0.86, Specificity 0.70 and Sensitivity 0.92.

4.1 Dataset

To train and evaluate the model, we selected the BraTS dataset [15]. The dataset consists of multimodal MRI images of glioblastoma (GBM/HGG) and lower-grade glioma (LGG) with pathologically confirmed diagnosis. Data are provided in train, validation, and test sets, but ground truth data are only available for train data. The training set consists of 260 HGG samples and 76 LGG samples. All BraTS multimodal scans are available in multiple clinical protocols as T1- and postcontrast T1-weighted, T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes. Segmented annotations are available for each volume. In the evaluation of the proposed method, we looked at several factors which influence the accuracy of the model. We evaluate the performance of the model with metrics such as Accuracy and F1 weighted score. Alongside the evaluation of the model performance, we focused on evaluating how the proposed loss function influenced the model in the way of choosing the right features in a decision process. To quantify this metric, we calculate the mean of \( \text{TumorLRPScore} \) for each class (LGG and HGG) in each experiment. The score was defined in the 3.6 section and generally tells how the model targets the tumor region.

4.2 Experiment 1

In table 1 below, we can see the results from the experiment in which we trained four identical models but the loss function was modified. The first model was trained with the Original Categorical cross-entropy loss function, this model is our baseline model without any changes in the training phase. The rest of the models are trained with the introduced loss function where TumorLRPScore is exponentiated to the power of 1, 2, and 3 as Penalization1/2/3. By increasing exponents, we test the penalization of the loss value with a higher power. This can be interpreted as defining the level of prioritization for features in the relevant area.

| Loss function | Accuracy | F1 score | LGG sc. | HGG sc. |
|---------------|----------|----------|---------|---------|
| Original (3)  | 0.72     | 0.71     | 0.26    | 0.22    |
| Penalization1 (4) | 0.79 | 0.79 | 0.36 | 0.36 |
| Penalization2 (5) | 0.72 | 0.70 | 0.43 | 0.61 |
| Penalization3 (6) | 0.74 | 0.74 | 0.46 | 0.65 |

The results showed a slight improvement in the accuracy of the models with the new loss function. An important observation is in the columns with the LRP score, which describes the attention of the model toward the tumor region. The results showed an increase in scores for both classes, with the gain being more pronounced for the HGG class. This could be an indication that our method helps the model better detect the features of the HGG class, which is more aggressive. We also observed the impact of the strength of the penalization, we see that higher power affects the fact of how much attention the model pays to the relevant regions in the input images.

4.3 Results visualisation

The presented results can also be interpreted in the following Fig. 5. Each row shows predictions of the four aforementioned models sequentially from the left. On the left, is the input image with the true label and the corresponding tumor segmentation mask followed by four LRP heatmaps, each for one of the trained models. Each heatmap is accompanied by pair of LRP scores and predicted labels drawn above.
The prediction examples in Fig. 5 represent the visual result of the proposed method application. In the heatmaps of the original model without changes to the loss function, we see that the LRP relevance is scattered throughout the brain. On the other hand, in the heatmaps where our method was used, we see an improvement in the intensity of the relevancies in the tumor regions, and also these relevancies are no longer scattered throughout the image. This is a practical projection of our proposed method.

4.4 Experiment 2

In the second experiment, we trained the state-of-the-art model from Subhas et al. [5] with the proposed explanation-driven training and our LRPLoss. Here, we observed the changes in the training process caused by our method compared to conventional training. The model was trained in 20 iterations and in each iteration, we captured the accuracy of the model on the validation data and the Mask LRP score. In the following Fig. 6, we see two plots describing the accuracy and Mask LRP scores for the model with the conventional loss function and our LRPLoss function configured as Penalization1 from the previous experiment.

From the graphs, we can see the impact of the proposed method. The main difference was gained in the first iterations of the training process, where our method helps the model achieve better accuracy in the earlier iterations. In later iterations, both models converged to similar accuracy. This behavior can be explained as an effect of our method where the LRPLoss function helps the model recognize faster the relevant features in the input image. This observation is supported by the evolution of the LRP mask score. In the first iteration, the score was significantly higher for the proposed method. Despite the decrease in scores in later iterations, the average Mask LRP score is maintained higher compared to the conventional training. We justify the convergence of accuracy and the decrease of Mask LRP scores as the dominance of crossentropy loss over penalization in later iterations, because when the model has enough information from the relevant part of the image, penalization becomes less important for prediction.

4.5 Results visualisation

The Fig. 7 shows the comparison of models during the training phase. The images show the relevant pixels obtained by LRP technique during the training iterations. We can see the practical impact of the proposed method where we can see that our method helps the model reveal relevant parts of the input image much faster than the conventional training.

5 Conclusion and future work

We have proposed a method in which the interpretation technique helps train the model in case we have any additional information related to the input data. We have shown that the LRP technique can be used not only to visualize the model’s decisions but also to improve the model. The method aims to force the model to pay more
Current DNN architectures usually require enormous amounts of data to learn from in order to achieve state-of-the-art performance. The proposed method of explanation-guided training can be used when we want to prefer some well-known features from the input, but we do not want to limit the model only to those features by segmentation or those areas that are scattered in the image. With explanation-guided training, we have reached faster convergence of the model. Current metrics of DNN solely focus on the model’s performance but it may lead to overfitting or learning inappropriate patterns. By prioritization of the features selected by domain experts, we create more robust and understandable models. Current DNN architectures usually require enormous amounts of data to learn from in order to achieve state-of-the-art performance.

In future work, we want to transfer and verify the proposed method for the next tasks of medical imaging using deep learning. As for instance for supporting a diagnostic process of Alzheimer’s disease, where attention is paid to anatomical parts of the brain and their features like by atrophy affected amygdala and hippocampus.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

1. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
2. Akkus Z, Ali I, Sedlák J, Agrawal JP, Parney IF, Giannini C, Erickson BJ (2017) Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from mr images using machine intelligence. J Dig Imaging 30(4):469–476
3. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fus 58:82–115
4. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10(7)
5. Banerjee S, Mitra S, Masulli F, Rovetta S (2019) Deep radiomics for brain tumor detection and classification from multi-sequence MRI. arXiv:1903.09240
6. Chetty G, Yamin M, White M (2022) A low resource 3d unet based deep learning model for medical image analysis. Int J Inf Technol 14(1):95–103
7. Cho HH, Park H (2017) Classification of low-grade and high-grade glioma using multi-modal image radiomics features. In: 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 3081–3084
8. Eberle O, Büttner J, Kräutli F, Müller K.R, Valleriani M, Montavon G (2020) Building and interpreting deep similarity models. arXiv preprint arXiv:2003.05431
9. Gunning D (2017) Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web, vol 2
10. Halder TK, Sarkar K, Mandal A, Sarkar S (2022) A novel histogram feature for brain tumor detection. Int J Inf Technol 14(4):1883–1892
11. Holzinger A, Biemann C, Pattichis C.S, Kell D.B (2017) What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923
12. Kalaiselvi T, Padmapriya S, Sriramakrishnan P, Somasundaram K (2020) Deriving tumor detection models using convolutional neural networks from MRI of human brain scans. Int J Inf Technol 12(2):403–408
13. Lipton ZC (2018) The mythis of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
14. Mehrrota R, Ansari M, Agrawal R, Anand R (2020) A transfer learning approach for AI-based classification of brain tumors. Mach Learn Appl 2:100003
15. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, West R et al (2014) The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans Med Imaging 34(10):1993–2024
16. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
17. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 65:211–222
18. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Dig Signal Process 73:1–15
19. Papernot N, McDaniel P (2018) Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. arXiv: 1803.04765
20. Petsiuk V, Das A, Saenko K (2018) Rise: Randomized input sampling for explanation of black-box models. arXiv:1806.07421
21. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk HV, Summers RM, Wiest R (2020) On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2(3):e190043
22. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
23. Samek W, Wiegand T, Müller KR (2018) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J ICT Discov Spec Issue 1 Impact Artif Intell AI Commun Netw Serv 1(1):39–48
24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
25. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034
26. Singh R, Agarwal BB (2022) An automated brain tumor classification in MR images using an enhanced convolutional neural network. Int J Inf Technol 1–10
27. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: the all convolutional net. arXiv:1412.6806
28. Sun J, Lapuschkin S, Samek W, Zhao Y, Cheung NM, Binder A (2020) Explanation-guided training for cross-domain few-shot classification. arXiv:2007.08790
29. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International conference on machine learning. PMLR, pp 3319–3328
30. Wu Y, Liu B, Wu W, Lin Y, Yang C, Wang M (2018) Grading glioma by radiomics with feature selection based on mutual information. J Ambient Intell Humaniz Comput 9(5):1671–1682
31. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, pp 818–833
32. Zurada J.M, Malinowski A, Cloete I (1994) Sensitivity analysis for minimization of input data dimension for feedforward neural network. In: Proceedings of IEEE international symposium on circuits and systems—ISCAS‘94, vol 6. IEEE, pp 447–450