Computer Methods and Programs in Biomedicine 243 (2024) 107867

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

SaRF: Saliency regularized feature learning improves MRI sequence classification

Suhang You [a], Roland Wiest [b,*], Mauricio Reyes [a,c]

[a] ARTORG, Graduate School for Cellular and Biomedical Research, University of Bern, Murtenstrasse 50, Bern, 3008, Switzerland
[b] Support Center of Advanced Neuroimaging, Institute of Diagnostic and Interventional Neuroradiology, University Hospital Bern, University of Bern, Freiburgstrasse 18, Bern, 3010, Switzerland
[c] Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse, Bern, 3010, Switzerland

ARTICLE INFO

Keywords:
MRI sequence classification
Interpretability
Saliency maps
Deep learning

ABSTRACT

Background and objective: Deep learning based medical image analysis technologies have the potential to greatly improve the workflow of neuro-radiologists dealing routinely with multi-sequence MRI. However, an essential step for current deep learning systems employing multi-sequence MRI is to ensure that their sequence type is correctly assigned. This requirement is not easily satisfied in clinical practice and is subjected to protocol and human-prone errors. Although deep learning models are promising for image-based sequence classification, robustness, and reliability issues limit their application to clinical practice.

Methods: In this paper, we propose a novel method that uses saliency information to guide the learning of features for sequence classification. The method uses two self-supervised loss terms to first enhance the distinctiveness among class-specific saliency maps and, secondly, to promote similarity between class-specific saliency maps and learned deep features.

Results: On a cohort of 2100 patient cases comprising six different MR sequences per case, our method shows an improvement in mean accuracy by 4.4% (from 0.935 to 0.976), mean AUC by 1.2% (from 0.9851 to 0.9968), and mean F1 score by 20.5% (from 0.767 to 0.924). Furthermore, based on feedback from an expert neuroradiologist, we show that the proposed approach improves the interpretability of trained models as well as their calibration with reduced expected calibration error (by 30.8%, from 0.065 to 0.045). The code will be made publicly available.

Conclusions: In this paper, the proposed method shows an improvement in accuracy, AUC, and F1 score, as well as improved calibration and interpretability of resulting saliency maps.

1. Introduction

For many years, MRI has played an important role in the diagnostic process that helps doctors and radiologists in their clinical workflow. In neuroimaging, MRI images serve to diagnose patients, follow up on their response to treatment, and assist in their prognosis, among other tasks. To cover different pathological or physiological conditions, different sequences are typically needed to provide different types of information. Hence, correct annotation of image sequences is a very important initial step of the image process pipeline in the clinics.

As more MRI scanners are introduced to medical centers, the amount of collected MRI images grows rapidly accordingly. The challenges of classifying collected sequences arise from a large amount of data. On top of that, scanners from different manufacturers have different naming formats as well as scanning protocols. Specifically, metadata from Digital Imaging and Communications in Medicine (DICOM) are not standardized throughout all medical centers, and each center might use different naming systems. For example, the “series description” (0008,103E) can vary from center to center due to local regulations, technician’s preferences, etc. [1]. In the clinical routine, accurate sequence annotation is challenged by the exhausting manual annotation required from professionals, which takes much effort and time. Automation is thus indeed highly requested, typically done via machine learning techniques. Previously, some methods have been proposed to tackle this issue using features extracted from DICOM metadata. [1] proposed to rely on the “series description” tag as ground truth for sequence clas-

* Corresponding author.
E-mail addresses: jaydensyou@outlook.com (S. You), roland.wiest@insel.ch (R. Wiest), mauricio.reyes@med.unibe.ch (M. Reyes).

https://doi.org/10.1016/j.cmpb.2023.107867
Received 26 May 2023; Received in revised form 11 October 2023; Accepted 15 October 2023
Available online 20 October 2023

sification. They report high-accuracy results (with accuracies ranging from 97.4% to 99.96%). However, as discussed, this tag has a large variability and can be very unreliable in clinical practice. [2] proposed to combine DICOM metadata and pixel information to train a deep learning sequence classification model. A few proposed methods are based on training deep learning models to classify image sequence types. [3] applied AlexNet [4] or GoogleNet [5], which reported accuracy levels ranging from 60.7% to 100%, while [6] proposed to use a modified Visual Geometry Group network (VGGNet [7]) accompanied by data augmentation, which reported the accuracy at 99%. Similarly, [8] utilized selected three-dimensional-cropped slabs of brain images during model training and leveraged test domain data to improve model performance (with an accuracy of 96.81%), and [9] only relied on single slice per volume to classify different sequences using ResNet18 [10] (with an accuracy of 79%), ResNet34 (with an accuracy of 81%), and ResNet50 (with an accuracy of 84%) networks.

In real-life applications, these methods can be trained and applied to internal datasets from a medical center (i.e., trained and tested on a local dataset). However, in MR-related deep learning applications, it is known that model performance drops when disparity exists between training and test datasets [11–13]. Hence, training models with multi-center datasets is a desired property of modern deep learning systems, which require large amounts of data, typically from public datasets. However, for neuroimaging applications, publicly available datasets commonly need to be skull-stripped for anonymization and data protection regulations [14]. For this reason, we postulate that a more realistic clinical scenario is to assess the performance of deep learning models, trained with skull-stripped brain images, and tested on original brain images. This follows the rationale that models trained on publicly available multi-center datasets aim at classifying image sequences at the very beginning of the image processing pipeline of a local center to enable further processing steps, including, for instance, skull-stripping. Another strategy could be running a sequence classification after skull-stripping. However, current skull-stripping methods are designed to work on either a specific sequence (typically T1-weighted) or have been trained to cover a few sequences, as is the case of [15] and likely will have issues on sequences not seen during training. [16] also concludes that skull-stripping methods need to be reliable regarding sequence variations, suggesting the dependency of these algorithms on the sequence type. In addition, the sequence classification followed by the skull-stripping scenario also emerged in our workflow as part of a multi-center project on multisequence brain MRI (The Swiss-First project [17]), where automated MRI sequence classification was requested due to the time-consuming nature of the task for the clinical collaborators of the project. We note that this is not a specific scenario but rather a typical one encountered in multicenter and multisequence clinical studies. Since such a more clinically-oriented scenario has not been explored in the literature, we extended the project's objectives to develop a solution presented in this study. The Swiss-First project uses acquisition protocols that are already more strict than daily clinical routine, yet the issue of MRI sequence classification still appeared. We believe this problem has been overlooked in the medical image computing community.

Another important challenge that has been rarely mentioned is inaccurately classifying T1-weighted MRI sequences (T1) and T1-post-Gadolinium (T1Gd) MRI sequences, which are very similar in appearance [18] (except for the areas where the contrast agent takes up), but have quite different objectives from a diagnostic and deep learning point of view (e.g., in brain tumors, the T1Gd sequence is the main driver in segmenting the active area of the tumor). Recently, in [9], the authors showed that T1Gd is easily misclassified as T1.

In summary, MRI sequence classification using deep learning has shown promising results. However, for clinical routine, model performance needs to be robust and highly accurate since any mistakes made at this initial point of the processing pipeline can have large detrimental effects on subsequent steps.

Saliency maps are one type of output yielded by the interpretability of models. Saliency maps are used to highlight which areas of an image are more important to a model. Beyond interpretation, saliency maps have also been used as an inductive bias during model training [19] and optimization of data augmentation policies [20]. In these works, saliency maps are used to enforce that saliency maps across different classes are as distinctive as possible.

Motivated by these works, we investigated the possibility of introducing saliency information during the training of a sequence classification model. We propose to do this via two self-supervised loss terms used during model training, such that the proposed saliency-driven loss terms act as a regularizer, enhancing the distinctiveness among class-specific saliency information, as well as the similarity between saliency information and deep features learned for the same class.

On a cohort of 2100 patient cases, comprising six different MR sequences per case, our result shows improved performance in the clinical scenario where sequence classification models are trained on skull-stripped brain images and tested on original brain images. At the sequence level, we also show the ability of the approach to drastically improve results on the hardest classification between T1 and T1Gd sequences. Furthermore, we show that the proposed approach improves the interpretability of trained models as well as their calibration.

In the following sections, we describe the proposed approach, termed SaRF, for Saliency Regularized Features, as well as the data and the implemented experimental setup.

2. Methods

In this section, we first introduce the MRI image sequences that are classified in this work, followed by a description of our proposed SaRF approach.

2.1. MRI image sequences

In this study, we focused on six MR sequences routinely used. In clinical practice, the first four types of sequences are used together for brain tumor diagnosis, which are T1-weighted images (T1), T2-weighted images (T2), T1-weighted images with Gadolinium contrast agent (T1Gd), and Fluid-attenuated inversion recovery (FLAIR) images. T1 images are widely used for brain anatomy. T2 images are widely used to detect image tissue anomalies such as brain tumors and lesions. Compared to T2 images, FLAIR images are able to suppress hyperintense signals from cerebrospinal fluids, providing better localization for anomalies. T1Gd images can highlight active tumoral areas, making it an essential sequence for a broad spectrum of disorders of the central nervous system (CNS) featuring a breakdown of the blood-brain-barrier (BBB), such as multiple sclerosis, stroke, infection, brain tumors, etc., for both detection and therapeutic guidance. In medical imaging computing-related challenges, such as the brain tumor segmentation challenge (BraTS) [21], and the Ischemic Stroke Lesion Segmentation (ISLES) [22,23] challenge, these four sequences are also used.

Apart from that, two other sequence types are also commonly used: diffusion-weighted imaging (DWI) and susceptibility-weighted imaging (SWI). DWI images help to distinguish benign and malignant tumors based on reduced diffusion or elevated diffusion. SWI images help to identify various compounds, such as blood, iron, and diamagnetic calcium, based on the magnetic susceptibility differences of these compounds. Fig. 3 shows examples of these six different sequences for one patient case.

2.2. Saliency-regularized features: SaRF

The overview of SaRF is shown in Fig. 1. During model training, class-specific saliency maps are calculated for each training image (e.g., using Deep Taylor Decomposition during a backward process) and for all potential sequence types (in this example, six) and aggregated to

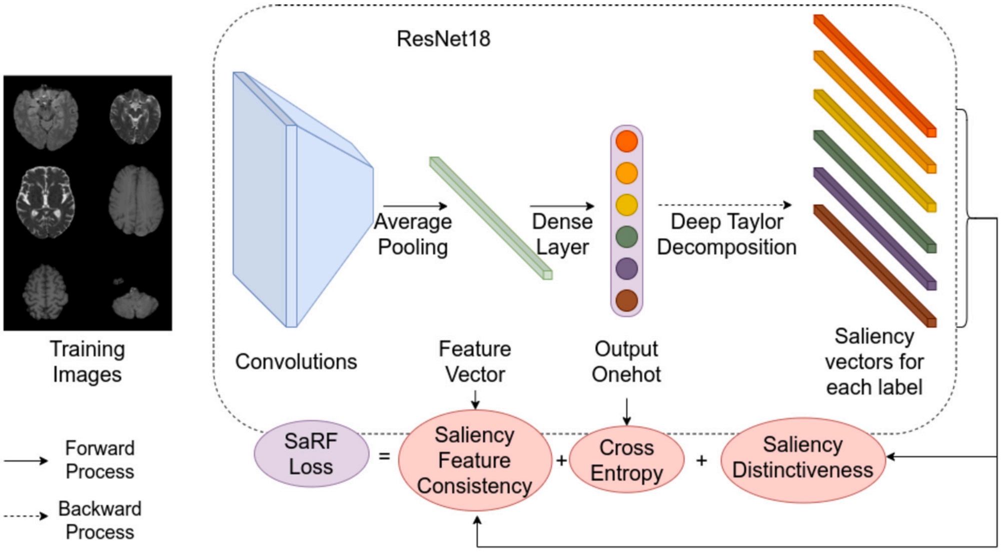

Fig. 1. Overview of the proposed SaRF approach. During model training, for each training image, class-specific saliency maps are calculated (e.g., using Deep Taylor Decomposition during a backward process) for all potential sequence types (in this example, six) and aggregated to yield a class-distinctiveness loss. Additionally, for each class, a feature vector from the second-to-last layer (forward pass) is extracted and compared to the corresponding class-specific saliency vector, to form a saliency-feature loss. These two terms are combined with the standard cross-entropy loss term to form the final loss function used for model training.

yield a class-distinctiveness loss. Additionally, for each class, a feature vector from the second-to-last layer (forward process) is extracted and compared to all (six) class-specific saliency vectors, to form a saliency-feature loss. These two terms are combined with the standard cross-entropy loss term to form the final loss function used during model training.

Given an input image, saliency maps show levels of pixel importance to the given task. High intensities in the saliency map reflect larger pixel attribution and impact on the task. In a classification task setting comprising K classes, a total of K class-specific saliency maps can be calculated. This is referred to as intra-sample saliency maps [19]. Based on this property, we aim to drive the training process to derive distinctive intra-sample saliency maps. Some interpretability approaches, such as Deep Taylor Decomposition (DTD) [24], enable calculating saliency maps at a specified layer. We take this advantage and calculate saliency maps to regularize the learned feature at that specified layer.

In general, the loss objective of the proposed SaRF is defined as the linear combination of a cross-entropy loss $\mathcal{L}_{ce}$, a saliency distinctiveness loss $\mathcal{L}_{sd}$, and a saliency-feature consistency loss $\mathcal{L}_{sf}$:

$$\mathcal{L}_{total}(I) = \mathcal{L}_{ce}(I) + \alpha \cdot \mathcal{L}_{sd}(I) + \beta \cdot \mathcal{L}_{sf}(I), \tag{1}$$

where $I \in \mathbb{R}^N$ is the input image, α and β are hyper-parameters that control the weight of each loss term. Below, we detail each saliency regularization loss term.

2.2.1. Saliency distinctiveness loss $\mathcal{L}_{sd}$

The proposed saliency distinctiveness loss $\mathcal{L}_{sd}$ aims at promoting the distinctiveness of class-specific saliency maps.

$$\mathcal{L}_{sd}(I) = \binom{K}{2}^{-1} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left| CosSim(S_i^L(I), S_j^L(I)) \right|, \tag{2}$$

where $CosSim(S_i^L(I), S_j^L(I))$ is the cosine similarity between the saliency of class i and j for input image I at layer L. $\binom{K}{2}^{-1}$ takes the mean of all calculated cosine similarity values from all combinations of class pairs for K classes where $\binom{}{}$ is the combination operator.

Our study uses DTD to calculate saliency information at the second-to-last layer. In the setting of DTD, for an input image $\mathbf{I} \in \mathrm{R}^N$, the prediction output $F(\mathbf{I})$ is decomposed backward layer by layer. In this process, each neuron is assigned a relevance score, and the assigning process is approximated by Taylor expansion. The collection of the relevance score for one layer is the saliency of this layer. This process starts at the output layer and eventually back-propagates to the input image space as shown in Fig. 3. During the process, the saliency of each layer can be calculated. This saliency information is a vector at the second-to-last layer, hence termed saliency vector. In section 4, we visualize the relevance score in the image space to show the effect of the proposed regularization terms.

2.2.2. Saliency-feature consistency loss $\mathcal{L}_{sf}$

The second regularization term of the proposed SaRF is the saliency-feature consistent loss $\mathcal{L}_{sf}$, which aims at enforcing the similarity between the calculated saliency vector and the learned feature for a given class.

$$\mathcal{L}_{sf}(I) = \frac{1}{K}\Big[\sum_{i=1, i\neq l}^{K} \left| CosSim(S_i^L(I), F_l^L(I)) \right| - \left| CosSim(S_l^L(I), F_l^L(I)) \right| \Big] \tag{3}$$

where $F_l^L(I)$ is the feature of input image I at layer L whose ground truth class is l. $CosSim(S_*^L(I), F_l^L(I))$ is the cosine similarity between the saliency $S_*^L(I)$ at layer L and the feature $F_l^L(I)$. The intuition of this loss is to enforce that learned features are similar to saliency maps computed for the ground truth class, while being orthogonal to the saliency vectors of the other classes. This intuition is enlightened by the findings in [25], which emphasizes that the explanation of a model should perfectly approximate the targeted model. Based on this perspective, we studied the possibility of enforcing the alignment between produced saliency maps and deep features of a model. In this way, the proposed saliency-feature consistency loss term of SaRF yields saliency maps that approximate learned features.

3. Experiments

3.1. Datasets

Our dataset comes from the Swiss-First study [17] aiming at performing early detection of epilepsy. Around 500 to 600 patients across

Table 1
Data source description. The Swiss-First study is within a multi-center project where there is no unified scanning protocol. The content provided is the subject number and the parameter range used during image acquisition. We provide the summary of the dataset in terms of magnetic field strength (1.5T or 3T), slice thickness range (in mm), spacing range (in mm), pixel size range (in mm × mm), scanning vendor manufacturer (Philips, Siemens or GE, in total 350 subjects), Gender (Female or Male, in total 350 subjects), and age range.

Sequence	DWI	FLAIR	SWI	T1	T1Gd	T2
Field Strength	95 (1.5T) 255 (3T)	89 (1.5T) 261 (3T)	104 (1.5T) 246 (3T)	70 (1.5T) 280 (3T)	95 (1.5T) 255 (3T)	66 (1.5T) 284 (3T)
Slice Thickness	1.5 ~ 5	0.7 ~ 5	1.2 ~ 14.4	0.9 ~ 6	0.7 ~ 6	1 ~ 5
Spacing	1.5 ~6.5	0.7 ~6	0.85 ~1.3	0.9 ~7.2	0.75 ~7.2	3 ~6.5
Pixel Size	0.55×0.55 ~ 1.81×1.81	0.43×0.43 ~ 1.3×1.3	0.29×0.29 ~ 0.94×0.94	0.24×0.24 ~ 1×1	0.24×0.24 ~ 1×1	0.21×0.21 ~ 0.45×0.45
Philips	12	12	8	15	14	24
Siemens	338	338	341	334	335	326
GE	0	0	1	1	1	0
Female	167	162	167	163	172	174
Male	183	188	183	187	178	176
Age	18~92	18~92	18~92	18~92	18~92	18~93

different centers in Switzerland are included in the study. In this study, we randomly selected 350 brain volumes for each sequence to have a balanced representation of sequence types, for a total of 2100 image volumes. The selected volumes contain both epileptic and healthy image volumes.

Table 1 describes the main characteristics of the dataset used for the study. For all brain images selected in this dataset, we collect the statistics of important scanning information. For example, the number of brain images scanned in 1.5 T and 3 T, the min slice thickness and the max slice thickness, etc.

3.2. Data split and pre-processing

The 350 image volumes per sequence were randomly split into 250 for model training and validation purposes and the remaining 100 image volumes were used to construct the test dataset. For the training and evaluation set (250 volumes), we apply 5-fold cross-validation and each fold consists of 200 image volumes for training and 50 image volumes for evaluation. The models trained in different folds will be tested on the test dataset.

During pre-processing, the selected brains were roughly aligned to the same origin and cropped in a similar range of 140 mm in the cranio-caudal direction to cover the brain tissue and skull.

To simulate the aforementioned scenario where training images stem from publicly available skull-stripped datasets, we use HD-BET [15] to perform skull-stripping on the training and validation datasets. Accordingly, we note that the test dataset was not skull-stripped. The test data in this scenario corresponds to the original (non-skull-stripped) brain images since such a classification system is meant to be used as the first step of a multi-sequence or multi-modal deep learning pipeline where knowing the sequences is needed to correctly order the inputs to the deep learning during inference time. We performed z-score normalization for each image volume and selected 7 slices at percentiles [20, 30, 40, 50, 60, 70, 80] in the Cranio-caudal direction, resulting in 1400 slices per class. For a 2D convolutional network, in total, we have 8400 slices for training. We performed this selection to ensure foreground presence on selected slices used for model training.

3.3. Model evaluation

As introduced in section 3.2, we perform 5-fold cross-validation. The model trained in each fold will select the model parameter with the smallest validation loss and then be evaluated on the test dataset. To predict the class of an input image volume, we make slice-wise predictions on the selected 7 slices, and predict the mean probability vectors, followed by selecting the maximum probability as the predicted class for the image volume. As classification metrics, we report accuracy, area under the curve (AUC) value from the receiver operating characteristic (ROC), and F1 score for each class. We also report mean accuracy (ACC), AUC, and F1 across all six classes for one test.

We also evaluated the calibration level of trained models, which is important to assess the reliability of the model's prediction confidence. Well-calibrated models match realistic predictions. If the model is poorly calibrated, the prediction value is not reliable. Poorly calibrated models can be over-confident or under-confident. To quantify the calibration of the model, the expected calibration error (ECE) [26] is typically used. We also visualize the calibration using reliability diagram [27]. We use the python and python library scikit-learn [28] for the metric calculations.

3.4. SaRF and ablations

As a baseline, we set $\alpha = 0$ and $\beta = 0$ to train a model, equivalent to training a model only with the cross-entropy loss. For the proposed SaRF, we set $\alpha = 5, \beta = 0.5$ to train the model. We also compare our result against focal loss, which we believe is a competitive alternative and used in medical image classification [29–32].

To investigate the effect of the regularization terms, we turned 'on' and 'off' the saliency distinctive loss and the saliency-feature consistency loss by setting $\alpha = 5/0$ or $\beta = 0.5/0$ when training models.

We further investigated how the α and β settings would affect the performance by conducting 64 combinations of α and β pairs taking values in the range [0, 0.2, 0.5, 1, 2, 5, 10, 20] when training models.

To investigate the generalizability of the proposed method, we also experiment with a different saliency method, GradCAM [33], to calculate the saliency information involved during the model's training. In addition, we experiment with focal loss [34] that replaces cross-entropy

Table 2

Classification results for each sequence type between the baseline and the proposed SaRF ($\alpha = 5, \beta = 0.5$), and average across all of them. SaRF yielded major improvements (> 10%) in classifying SWI, T1, and T1Gd. The improvement in classifying the T1 sequence is the largest across all six sequences. The best results in F1 score with major improvement (> 10%) are shown in bold, and the results in italics indicate worsened results.

Sequence	Methods	ACC	AUC	F1
DWI	Baseline	0.973 ± 0.032	0.9996 ± 0.0008	0.929 ± 0.080
	SaRF	0.996 ± 0.002	0.9999 ± 0.0001	0.989 ± 0.006
FLAIR	Baseline	0.981 ± 0.009	0.9817 ± 0.0117	0.939 ± 0.030
	SaRF	0.991 ± 0.003	0.9930 ± 0.0044	0.972 ± 0.008
SWI	Baseline	0.943 ± 0.045	0.9983 ± 0.0030	0.862 ± 0.094
	SaRF	0.994 ± 0.005	0.9999 ± 0.0002	**0.982 ± 0.016**
T1	Baseline	0.855 ± 0.018	0.9696 ± 0.0218	0.226 ± 0.163
	SaRF	0.944 ± 0.037	0.9970 ± 0.0031	**0.782 ± 0.162**
T1Gd	Baseline	0.896 ± 0.018	0.9612 ± 0.0293	0.726 ± 0.068
	SaRF	0.945 ± 0.029	0.9907 ± 0.0084	**0.857 ± 0.069**
T2	Baseline	0.962 ± 0.071	0.9999 ± 0.0000	0.919 ± 0.014
	SaRF	0.985 ± 0.024	*0.9999 ± 0.0001*	0.961 ± 0.060
ALL	Baseline	0.935 ± 0.019	0.9851 ± 0.0103	0.767 ± 0.068
	SaRF	0.976 ± 0.014	0.9968 ± 0.0020	**0.924 ± 0.046**

and apply the proposed SaRF loss terms. We use the same parameter setting for these two experiments as the proposed SaRF, with $\alpha = 5, \beta = 0.5$ to train the model.

3.5. Faithfulness study on learned features

To quantitatively analyze if the proposed SaRF improves the interpretability of learned features, we followed the QUANTUS [35] framework and performed the faithfulness study. In QUANTUS, the faithfulness study is defined to quantify to what extent explanations follow the predictive behavior of the model, asserting that more important features affect model decisions more strongly. It was originally defined to compare different interpretability methods, but it also applies to this work, which assumes that improved feature learning leads to more important features being highlighted in DTD saliency maps. For faithfulness, we adopted the Remove and Debias (ROAD) approach [36] with the 'Most Relevant First' order [37], which measures the accuracy of a model on a test set in an iterative process by removing the k-most important pixels (as yielded by the corresponding saliency map). The pixel intensity is replaced by an imputed value from neighbors with randomly sampled linear noise during the removal. A saliency map is more faithful if the drop in model performance is higher when perturbing the most important pixels in the MRI image. Following the definition of faithfulness, we can quantify how the proposed SaRF inductive bias yields improved learned features compared to the baseline model. We applied the faithfulness test to all test images and predicted the volume classes at different removal levels. The result is evaluated with the F1 score since it is the most sensitive metric.

3.6. Training details

We used ResNet18 as the backbone architecture, with minor adaptations to work with medical images and saliency calculation. During training, we use RMSprop optimizer at the learning rate of 10^{-5}. The weight decay was set to 10^{-4} with momentum set to 0.9. We set the number of training epochs to 200 to ensure convergence of the loss. After training, we selected the saved model yielding the best validation loss at the end of each epoch for testing. All the experiments were executed with Pytorch running on Nvidia Geforce GTX 1080ti graphic cards. The code will be available via GitHub https://github.com/yousuhang/SaRF.

4. Results

4.1. Comparison of SaRF and the baseline

General results are shown in Table 2. The results in bold show major improvement (>10%), and the results in italics indicate worsened results. The mean and standard deviation for each metric was calculated from 5-fold cross-validation on the test dataset. The results of both baseline and SaRF have values close to 1 in ACC and AUC. SaRF achieves improvements across the board for ACC and AUC. The major improvements are observed for the F1 score, with improvements of 20.5% (from 0.767 to 0.924), particularly for the T1 sequence, going from 0.226 to 0.782. These results demonstrate the ability of the proposed method to yield improved classification results, especially for the T1 sequence, which largely lacks accuracy for the baseline model.

We also plotted the average confusion matrix for the baseline and proposed SaRF, consisting of an average across 5-fold test results. In each confusion matrix plot of Fig. 2, rows correspond to the predicted classes, and the columns correspond to the ground truth classes. Results were normalized with respect to the ground truth class, where the summation of each row is 1. For example, in the confusion matrix for the baseline model, the T1 classification (row T1), 13.6% of predictions are correct, while 41% of T1 test sequences were classified as T1Gd by this model, and 21.8% T1 test sequences were classified as SWI. The confusion matrix shows that the major improvement occurs for the T1 and T1Gd sequences, wherein current approaches yield large classification errors. With the proposed SaRF approach, the improvement is remarkable, from 0.136 to 0.668 on T1, from 0.84 to 0.96 on T1Gd, and from 0.886 to 0.946 on FLAIR, while not sacrificing accuracy on other sequences.

To investigate the interpretability of trained models, we compared the saliency maps of SaRF and baseline models for test cases. In Fig. 3, we present one exemplary slice per sequence for one subject. We selected the slice containing a large area of non-brain tissue to investigate how these regions are used for each model since the training data is skull-stripped. In Fig. 3, each row shows the input slice on the left side and saliency maps on the right side divided into two smaller rows. The top row shows the saliency maps for the baseline model, whereas the bottom rows show saliency maps for models trained with SaRF. For example, in the row DWI and the column DWI, the upper image is the

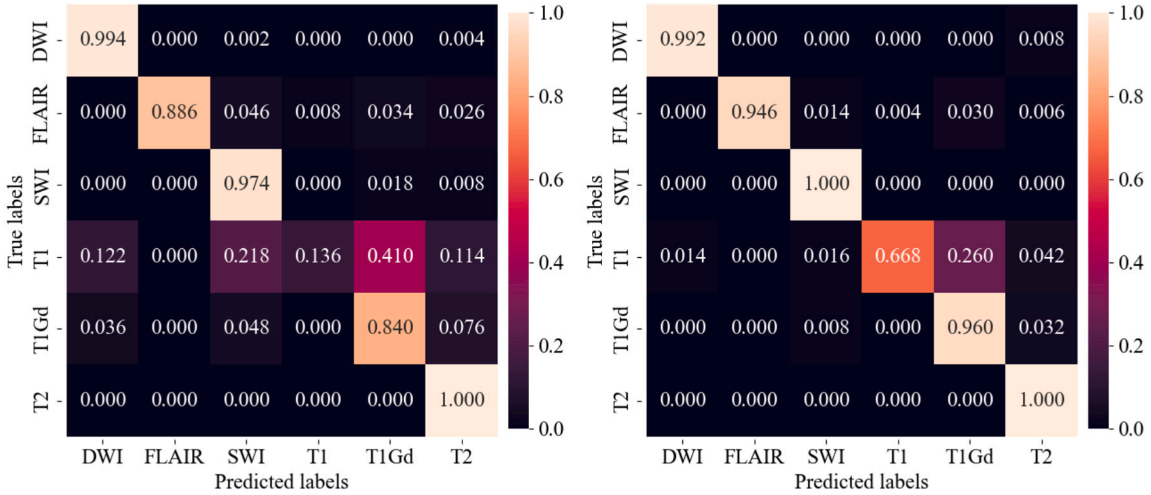

Fig. 2. Mean Confusion Matrix of the baseline in the left and SaRF $\alpha = 5, \beta = 0.5$ in the right. The horizontal is the predicted class, and the vertical is the ground truth class. The values are normalized horizontally for each ground truth class, where the summation of each row is 1.

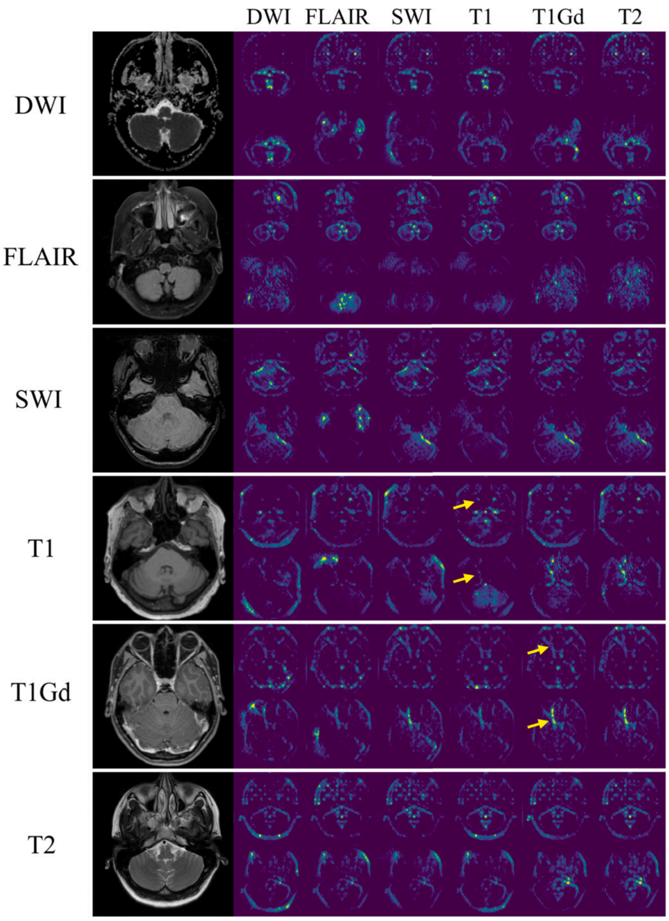

Fig. 3. Input image in the left and calculated saliency maps for each class in the right. For each input image, the top row are the saliency maps for the baseline, and the bottom row are the saliency maps for SaRF $\alpha = 5, \beta = 0.5$. Input images are examples from the test dataset. On the right, each column are the saliency maps calculated for the denoted class. For example, in the row DWI and the column DWI, the upper image is the saliency map calculated for the input DWI image for class DWI for the baseline method, and the lower image is the saliency map calculated for the same input and the same class for the proposed SaRF. The yellow arrows point out the muscle tissue linked to the eyes as suggested by the neuroradiologist.

Table 3

F1 scores of the classification results for each sequence type and average across all six sequences. Each column shows the result of the model evaluated on the test set in a 5-fold cross-validation. From left to right are the baseline method, the method only uses focal loss as the loss function, the proposed SaRF ($\alpha = 5, \beta = 0.5$), SaRF using GradCAM in saliency calculation ($\alpha = 5, \beta = 0.5$), SaRF using focal loss to replace the cross-entropy ($\alpha = 5, \beta = 0.5$). The best results in the 'ALL' row with major improvement (> 10%) are shown in bold.

Sequence	Baseline	Focal Loss	SaRF	SaRF (GradCAM)	SaRF (Focal Loss)
DWI	0.929 ± 0.080	0.936 ± 0.053	0.989 ± 0.006	0.962 ± 0.060	0.992 ± 0.010
FLAIR	0.939 ± 0.030	0.947 ± 0.026	0.972 ± 0.008	0.926 ± 0.052	0.970 ± 0.007
SWI	0.862 ± 0.094	0.805 ± 0.112	0.982 ± 0.016	0.950 ± 0.032	0.991 ± 0.007
T1	0.226 ± 0.163	0.222 ± 0.133	0.782 ± 0.162	0.389 ± 0.213	0.799 ± 0.060
T1Gd	0.726 ± 0.068	0.671 ± 0.126	0.857 ± 0.069	0.706 ± 0.061	0.851 ± 0.032
T2	0.919 ± 0.014	0.957 ± 0.042	0.961 ± 0.060	0.969 ± 0.227	0.943 ± 0.070
ALL	0.767 ± 0.068	0.756 ± 0.071	**0.924 ± 0.046**	0.817 ± 0.057	**0.924 ± 0.027**

saliency map calculated for the input DWI image for class DWI for the baseline method, and the lower image is the saliency map calculated for the same input and the same class for the proposed SaRF. A qualitative assessment shows that: 1) SaRF shows more distinctive saliency maps across different class. For example, for the FLAIR sequence, the saliency maps for the baseline model are similar among class, whereas they are visually more distinctive for SaRF. 2) SaRF can highlight some anatomically meaningful areas for neuroradiologists when classifying sequences. One example is shown in Fig. 3. The yellow arrows indicate that the muscle tissue linked to the eyes is much more highlighted in SaRF compared to the baseline in the T1Gd sequence (T1Gd-T1GD column-row), while in T1, conversely not highlighted (T1-T1 column-row). This agrees more with the criteria in manual classification between T1 and T1Gd, whereas in the baseline method, the muscle tissue is not specially highlighted in the T1Gd-T1GD column-row but more highlighted in T1-T1 column-row (the muscle linked to the right eye). 3) The skull and non-brain tissues contribute to the classification process for both the baseline and regularized models.

4.2. Faithfulness test between SaRF and the baseline

The result of the faithfulness test comparing SaRF and the baseline is shown in Fig. 5-(a). Since SaRF and the baseline have different F1 scores, as shown in Table 2 ('ALL' row), we measure the decrease of F1 score when different percentages of removal (x-axis) are applied to the test dataset. The decreased value (y-axis) is calculated by subtracting the mean F1 scores at different removal percentages from the F1 score without removal. Therefore, the *larger* decreased value is *better*. In the initial 10% most relevant pixels removal, these two methods show no apparent difference in F1 decrease. As the removal process continues, the proposed SaRF shows a larger decrease in F1 and, eventually, a gap (around 0.1) between the two methods. This shows that the proposed SaRF leads to more faithful learned features.

4.3. Ablations on comparing different methods

The results of ablations comparing F1 scores of different methods are shown in Table 3. Other metrics (ACC and AUC) results are presented in Appendix A. From left to right are the baseline method, focal loss as the single loss function, the proposed SaRF ($\alpha = 5, \beta = 0.5$), SaRF using GradCAM in saliency calculation ($\alpha = 5, \beta = 0.5$), SaRF using focal loss to replace the cross-entropy ($\alpha = 5, \beta = 0.5$). The mean and standard deviation are calculated from 5-fold cross-validation on the test dataset. Comparing the baseline and the focal loss method, they have similar performance, which shows that the selected baseline is solid, as used in former works [3,6,8,9]. The SaRF calculated with GradCAM during saliency calculation has an improvement of 6.5% (from 0.767 to 0.817) for the F1 score but is lower than the proposed SaRF using DTD (20.5% improvement). A potential reason for this difference is that the calculation of GradCAM is different from DTD, and the gradient scale of each saliency loss term differs from the proposed SaRF's saliency loss terms. Therefore, the selected parameter setting ($\alpha = 5, \beta = 0.5$) might not be optimal for SaRF with GradCAM. The SaRF replacing cross-entropy with focal loss during model training has an improvement of 20.5% (from 0.767 to 0.924) for the F1 score, the same as the proposed SaRF (with DTD using cross-entropy loss). Both modified SaRF methods show major improvements in T1 sequence classification, going from F1 metric values of 0.226 to 0.389 and from 0.226 to 0.799, respectively. These results demonstrate the ability of the proposed SaRF loss terms to yield improved classification results also in the case of using a different saliency calculation method (i.e., SaRF with GradCAM) and the case applying to a different loss function (SaRF with focal loss). The ablation experiment comparing different methods shows that our proposed SaRF can be easily applied to different losses, and the SaRF loss terms can be calculated using different saliency methods.

4.4. Ablations on parameters of each saliency regularization term

The results of the ablation study analyzing the impact of each loss term of SaRF are shown in Fig. 4, using the F1 score (other metrics are presented in Appendix A). In Fig. 4, the baseline, SaRF, ablation $\alpha = 5, \beta = 0$, and ablation $\alpha = 0, \beta = 0.5$ are in blue, red, orange, and green, respectively. The major improvements occur for the sequences T1 and T1Gd. The results show that both the saliency-feature consistency loss term and the saliency distinctiveness loss term yield distinctive performance improvements except for the T2 sequence when applied separately. Combining the two loss terms, the performance is further improved for each classification compared to solely using each loss term.

We also compared the calibration levels of the trained model among the four studied ablation settings. Shown in Fig. 5-(b), the added regularization yields a reduction of ECE (lower is better). The results of the ablation study on other hyper-parameter settings for the saliency distinctiveness and saliency-feature consistency are shown in Fig. 5-(c). Other metrics results are presented in Appendix A (Fig. 6, Table 4 and Table 5). Comparing the variation of different parameter settings, we observe that the impact of the saliency-feature consistency loss is larger than the impact of the saliency distinctiveness loss term.

5. Discussion

Sequence classification is essential as a building block in deep learning imaging pipelines requiring multi-sequence inputs. Correct identification of sequence type is hence essential to ensure the correct utilization of imaging information by a trained deep learning model. Although this information can be included in DICOM metadata, in practice, there is large heterogeneity in how this information is encoded.

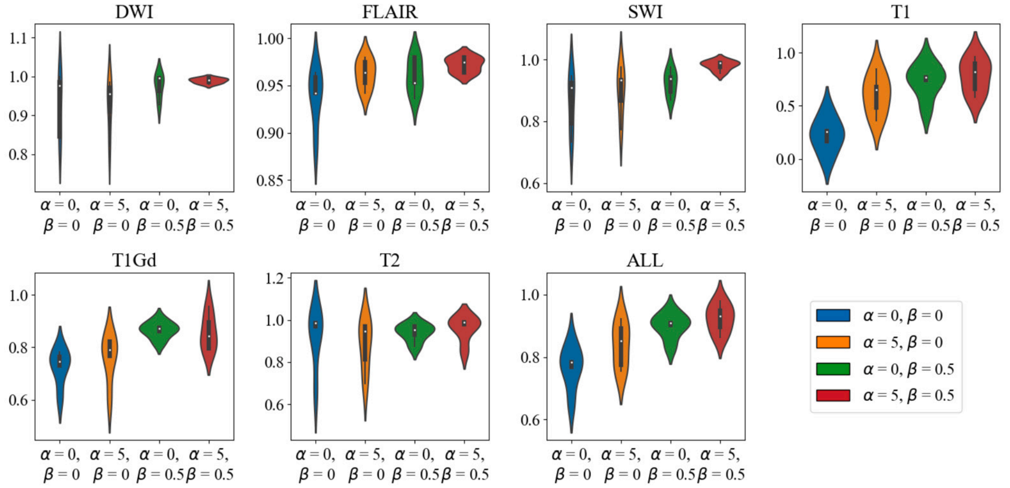

Fig. 4. Ablation study on the importance of each loss term of SaRF. Violin plot of test F1 scores of each setting for all cross-validation folds. The major improvements occur for the sequences T1 and T1Gd.

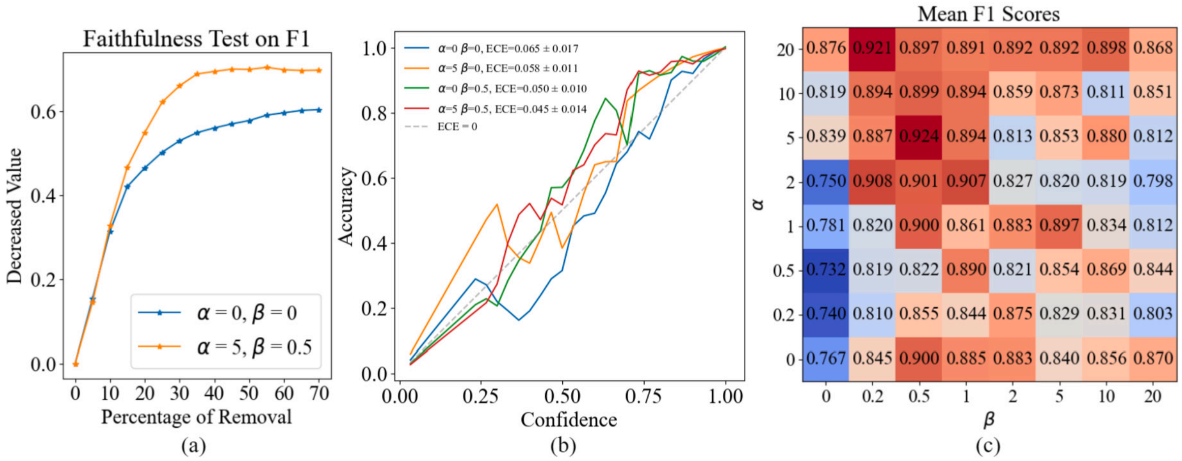

Fig. 5. (a) Faithfulness test on F1 score between the proposed SaRF (orange) and the baseline (blue). The x-axis is the percentage of the most salient pixels (highest intensity yielded in saliency maps for unmodified test images) removed from the input test images for prediction. The y-axis is the decreased value between the mean predicted F1 scores of the modified images and the mean predicted F1 scores of the unmodified test images, i.e., the F1 scores of the 'ALL' row in Table 2. *Higher* values are *better*. (b) The reliability diagram of models trained with four different α, β settings. Blue: Baseline, Red: SaRF, Orange: ablation $\alpha = 5, \beta = 0$, Green: ablation $\alpha = 0, \beta = 0.5$. Respectively, the ECE values are 0.065 ± 0.017, 0.058 ± 0.011, 0.05 ± 0.01, and 0.045 ± 0.014. The gray dash diagonal line represents a perfect calibration (ECE = 0). *Lower* ECE values are *better*. (c) Mean F1 scores for different hyper-parameter α, β pairs used during training. The mean F1 scores are tested and averaged across all cross-validation folds of trained models.

For example, a T1Gd (for gadolinium) sequence can be termed differently as "T1c", "T1ce", "T1-post-GAD", etc. Furthermore, it is also common that this information is not stored under the same tag values by different vendors or software tools. Due to this and other reasons, image-based sequence classification from imaging information provides an alternative mechanism to be incorporated in a hybrid fashion with DICOM-based sequence classification.

In this study, we propose to use interpretability information to guide the learning process of a sequence classification model. This is based on the strong link between model performance and model interpretability, as discussed in [38], and as reported as well for other tasks in [19,20]. One of the novel loss terms proposed in this study considers the similarity between the latent representation of saliency maps and the deep features. The rationale aligns with recent findings in [25], which suggests that a proper explanation of a model should align with its deep features. Although our faithfulness study shows an improvement in the alignment, the clinical meaning of such learned features still requires verification if the more highlighted features are more clinically significant. To address this issue, we consulted with a senior neuroradiologist to evaluate saliency maps from both the baseline and the proposed SaRF results. We presented to the neuroradiologist six randomly chosen sequences from different patient cases. On each sequence, we displayed six class-specific saliency maps (similar to those shown in Fig. 3. We blinded the methods and asked the expert to select per sequence, which saliency maps he would prefer, and provide justification. The expert commented that specifying brain regions that are sequence-specific is a difficult task, but he was able to find a few patterns he considered relevant to guide his selection. For example, the expert pointed out a checkerboard effect in the T2 sequence and believed that the method was attending to this area due to the inhomogeneity of magnetic field strength. This could be attributed to a potential shortcut learning of the model, while our regularized training seems to be resistant to this effect. However, even for the neuroradiologist, it is not known what the correct regions of interest per sequence should be (i.e., ground-truth saliency regions per sequence).

In the ablations study in section 4.3, we show that our method can be easily extended by changing the saliency calculation method (as the results with GradCAM) or by jointly working with other types of losses (as the results with the focal loss) and still achieve improved performance in the classification task. Beyond that, we also tested with ResNet34 and ResNet50 as backbone models, following [9]. Results showed that the proposed SaRF approach leads to significant improvements for both ResNet34 (from 0.747 to 0.852 in F1 scores) and ResNet50 (from 0.808 to 0.844 in F1 scores), suggesting the generalization capabilities of SaRF for different CNN architectures.

Some limitations are worth mentioning. In this study, we only include six types of sequences from the Swiss-First project. In other clinical settings, other types of sequences, such as perfusion-weighted imaging sequences, should also be included in future works. During the pre-processing of brain skull-striping, without access to the ground truth of skull-stripped brains from radiologists, we used HD-BET, which was originally trained with four types of sequences (T1, T1Gd, T2, and FLAIR). The tool might not perfectly skull-strip DWI and SWI sequences. Compared to the baseline with only cross-entropy loss, the proposed SaRF with saliency calculation increases the computation time by more than 50% during model training. As discussed above, the neuroradiologist's qualitative evaluation of saliency maps is limited by the number of cases and limited understanding of the connection between saliency maps and clinical decisions (ground-truth saliency regions per sequence).

In the future, we would like to investigate how to better characterize and define what regions of interest should mostly drive a sequence classification model, as well as to investigate how changes on saliency maps are quantitatively connected to performance improvements. Future work also includes conducting a larger qualitative analysis with clinical experts and extending the study to other MR sequences, including diffusion and perfusion MRI. Beyond classification, we think the approach can be extended to segmentation problems, provided interpretability saliency maps for segmentation tasks are available. For a practical challenge that might occur when a new imaging sequence type is introduced in the clinics, an interesting further venue of research is to combine the present approach with generalized zero-shot learning, as proposed in [39]. Beyond CNN, it would be interesting to extend the proposed to other types of architectures, such as vision transformers [40].

6. Conclusion

In this study, we proposed SaRF, a novel method that introduces saliency information via two self-supervised loss terms during the training of a deep learning classification model. The saliency distinctiveness loss enhances the distinctiveness among class-specific saliency maps, and the saliency-feature consistency loss enhances the similarity between saliency maps and corresponding learned deep features for the same class. The proposed SaRF shows an improvement in terms of accuracy, AUC, and F1 score, as well as improved calibration and interpretability of resulting saliency maps.

Ethical approval

The data used in this work is approved by the Swiss-First Study project.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgement

This work resulted from the research project funded by the Swiss Personalized Health Network (SPHN) initiative and supported by the Swiss National Science Foundation under grant number CRSII5_180365 (The Swiss-First Study).

Appendix A. Meam ACC & AUC heat-map and ACC & AUC comparison among different methods

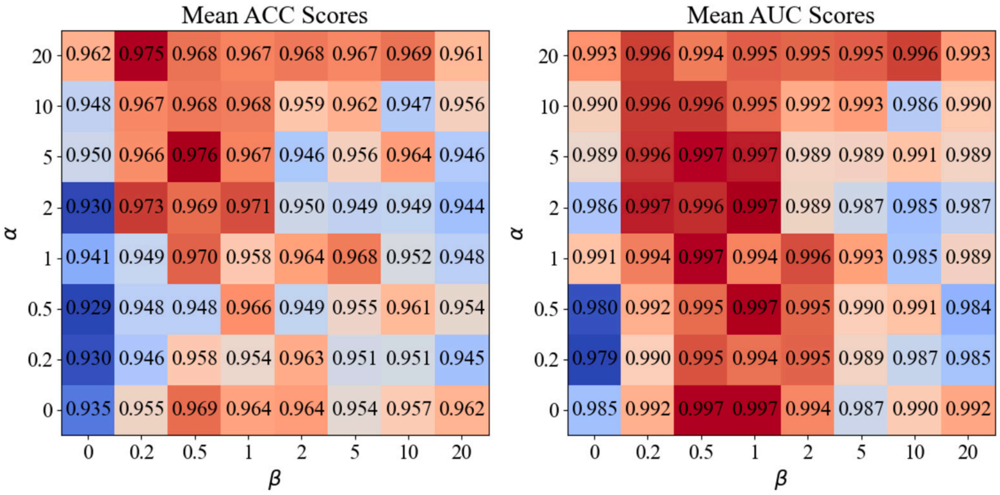

Fig. 6. Mean ACC heat-map (left) and Mean AUC heat-map (right) for different hyper-parameter α, β pairs used during training. The mean AUCs are tested and averaged across five folds.

Table 4
ACC scores of the classification results for each sequence type and average across all six sequences. Each column represents the method used in a 5-fold cross-validation. From left to right are the baseline method, the method only uses focal loss as the loss function, the proposed SaRF ($\alpha = 5, \beta = 0.5$), SaRF using GradCAM in saliency calculation ($\alpha = 5, \beta = 0.5$), SaRF using focal loss to replace the cross-entropy ($\alpha = 5, \beta = 0.5$).

Sequence	Baseline	Focal Loss	SaRF	SaRF (GradCAM)	SaRF (Focal Loss)
DWI	0.973 ± 0.032	0.977 ± 0.020	0.996 ± 0.002	0.986 ± 0.023	0.997 ± 0.003
FLAIR	0.981 ± 0.009	0.983 ± 0.008	0.991 ± 0.003	0.978 ± 0.014	0.990 ± 0.002
SWI	0.943 ± 0.045	0.912 ± 0.068	0.994 ± 0.005	0.983 ± 0.011	0.997 ± 0.002
T1	0.855 ± 0.018	0.855 ± 0.014	0.944 ± 0.037	0.877 ± 0.029	0.945 ± 0.014
T1Gd	0.896 ± 0.018	0.886 ± 0.010	0.945 ± 0.029	0.866 ± 0.039	0.945 ± 0.011
T2	0.962 ± 0.071	0.984 ± 0.016	0.985 ± 0.024	0.989 ± 0.008	0.978 ± 0.029
ALL	0.935 ± 0.019	0.933 ± 0.019	0.976 ± 0.014	0.946 ± 0.015	0.975 ± 0.009

Table 5
AUC scores of the classification results for each sequence type and average across all six sequences. Each column represents the method used in a 5-fold cross-validation. From left to right are the baseline method, the method only uses focal loss as the loss function, the proposed SaRF ($\alpha = 5, \beta = 0.5$), SaRF using GradCAM in saliency calculation ($\alpha = 5, \beta = 0.5$), SaRF using focal loss to replace the cross-entropy ($\alpha = 5, \beta = 0.5$).

Sequence	Baseline	Focal Loss	SaRF	SaRF (GradCAM)	SaRF (Focal Loss)
DWI	0.9996 ± 0.0008	0.9982 ± 0.0038	0.9999 ± 0.0001	0.9996 ± 0.0005	0.9999 ± 0.0000
FLAIR	0.9817 ± 0.0117	0.9897 ± 0.0072	0.9930 ± 0.0044	0.9900 ± 0.0003	0.9962 ± 0.0030
SWI	0.9983 ± 0.0030	0.9999 ± 0.0000	0.9999 ± 0.0002	0.9998 ± 0.0003	0.9999 ± 0.0000
T1	0.9696 ± 0.0218	0.9867 ± 0.0081	0.9970 ± 0.0031	0.9799 ± 0.0148	0.9973 ± 0.0023
T1Gd	0.9612 ± 0.0293	0.9589 ± 0.0184	0.9907 ± 0.0084	0.9621 ± 0.0300	0.9922 ± 0.0058
T2	0.9999 ± 0.0000	0.9999 ± 0.0000	0.9999 ± 0.0001	0.9999 ± 0.0002	0.9999 ± 0.0001
ALL	0.9851 ± 0.0103	0.9889 ± 0.0053	0.9968 ± 0.0020	0.9886 ± 0.0072	0.9976 ± 0.0015

References

[1] R. Gauriau, C. Bridge, L. Chen, F. Kitamura, N.A. Tenenholtz, J.E. Kirsch, K.P. Andriole, M.H. Michalski, B.C. Bizzo, Using dicom metadata for radiological image series categorization: a feasibility study on large clinical brain mri datasets, J. Digit. Imag. 33 (3) (2020) 747–762.

[2] J. Cluceru, J.M. Lupo, Y. Interian, R. Bove, J.C. Crane, Improving the automatic classification of brain mri acquisition contrast with machine learning, J. Digit. Imag. (2022) 1–17.

[3] T. Noguchi, D. Higa, T. Asada, Y. Kawata, A. Machitori, Y. Shida, T. Okafuji, K. Yokoyama, F. Uchiyama, T. Tajima, Artificial intelligence using neural network architecture for radiology (ainnar): classification of mr imaging sequences, Japan. J. Radiol. 36 (12) (2018) 691–697.

[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[6] S. Ranjbar, K.W. Singleton, P.R. Jackson, C.R. Rickertsen, S.A. Whitmire, K.R. Clark-Swanson, J.R. Mitchell, K.R. Swanson, L.S. Hu, A deep convolutional neural network for annotation of magnetic resonance imaging sequence type, J. Digit. Imag. 33 (2) (2020) 439–446.

[7] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

[8] J.P.V. de Mello, T.M. Paixão, R. Berriel, M. Reyes, C. Badue, A.F. De Souza, T. Oliveira-Santos, Deep learning-based type identification of volumetric mri sequences, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 1–8.

[9] N. Braeker, C. Schmitz, N. Wagner, B.J. Stanicki, C. Schröder, F. Ehret, C. Fürweger, D.R. Zwahlen, R. Förster, A. Muacevic, et al., Classifying the acquisition sequence for brain MRIs using neural networks on single slices, Cureus 14 (2) (2022) e22435, https://doi.org/10.7759/cureus.22435, publisher: Cureus.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, preprint, arXiv:1512.03385.

[11] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, 2006, https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

[12] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, E. Burnaev, Domain shift in computer vision models for mri data analysis: an overview, in: Thirteenth International Conference on Machine Vision, vol. 11605, SPIE, 2021, pp. 126–133.

[13] B. Glocker, R. Robinson, D.C. Castro, Q. Dou, E. Konukoglu, Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects, preprint, arXiv:1910.04597.

[14] T.E. Nichols, S. Das, S.B. Eickhoff, A.C. Evans, T. Glatard, M. Hanke, N. Kriegeskorte, M.P. Milham, R.A. Poldrack, J.-B. Poline, et al., Best practices in data analysis and sharing in neuroimaging using mri, Nat. Neurosci. 20 (3) (2017) 299–303.

[15] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al., Automated brain extraction of multisequence mri using artificial neural networks, Hum. Brain Mapp. 40 (17) (2019) 4952–4964.

[16] P. Kalavathi, V.S. Prasath, Methods on skull stripping of mri head scan images—a review, J. Digit. Imag. 29 (2016) 365–379.

[17] B.Z. Jin, P. De Stefano, V. Petroulia, C. Rummel, C. Kiefer, M. Reyes, K. Schindler, P. van Mierlo, M. Seeck, R. Wiest, Diagnosis of epilepsy after first seizure. introducing the swiss first study, Clin. Transl. Neurosci. 4 (2) (2020) 13.

[18] K. Usman, K. Rajpoot, Brain tumor classification from multi-modality mri using wavelets and machine learning, Pattern Anal. Appl. 20 (2017) 871–881.

[19] D. Mahapatra, A. Poellinger, M. Reyes, Interpretability-guided inductive bias for deep learning based medical image, Med. Image Anal. 81 (2022) 102551.

[20] S. You, M. Reyes, Sagtta: saliency guided test time augmentation for medical image segmentation across vendor domain shift, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–4.

[21] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (brats), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[22] O. Maier, B.H. Menze, J. Von der Gablentz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al., Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri, Med. Image Anal. 35 (2017) 250–269.

[23] M.R. Hernandez Petzsche, E. de la Rosa, U. Hanning, R. Wiest, W. Valenzuela, M. Reyes, M. Meyer, S.-L. Liew, F. Kofler, I. Ezhov, et al., Isles 2022: a multi-center magnetic resonance imaging stroke lesion segmentation dataset, Sci. Data 9 (1) (2022) 762.

[24] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognit. 65 (2017) 211–222.

[25] T. Han, S. Srinivas, H. Lakkaraju, Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations, preprint, arXiv:2206.01254.

[26] M.P. Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, 2015, pp. 2901–2907.

[27] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 625–632.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[29] Q. Lou, Y. Li, Y. Qian, F. Lu, J. Ma, Mammogram classification based on a novel convolutional neural network with efficient channel attention, Comput. Biol. Med. 150 (2022) 106082.

[30] S.R. Ahmed, A. Lemay, K.V. Hoebel, J. Kalpathy-Cramer, Focal loss improves repeatability of deep learning models, in: Medical Imaging with Deep Learning, 2022.

[31] G.S. Tran, T.P. Nghiem, V.T. Nguyen, C.M. Luong, J.-C. Burie, et al., Improving accuracy of lung nodule classification using deep learning with focal loss, J. Healthcare Eng. (2019).

[32] Z. Qiao, A. Bae, L.M. Glass, C. Xiao, J. Sun, Flannel (focal loss based neural network ensemble) for covid-19 detection, J. Am. Med. Inform. Assoc. 28 (3) (2021) 444–452.

[33] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[35] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M.M.-C. Höhne, Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations and beyond, J. Mach. Learn. Res. 24 (34) (2023) 1–11.

[36] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods, preprint, arXiv:2202.00449.

[37] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, A. Preece, Sanity checks for saliency metrics, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 6021–6029.

[38] S. You, M. Reyes, Influence of contrast and texture based image modifications on the performance and attention shift of u-net models for brain tissue segmentation, Front. Neuroimag. 1 (2022) 1012639, https://doi.org/10.3389/fnimg.2022.1012639.

[39] D. Mahapatra, Z. Ge, M. Reyes, Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps, IEEE Trans. Med. Imaging 41 (9) (2022) 2443–2456.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, preprint, arXiv:2010.11929.