Journal of Neuroscience Methods 353 (2021) 109098

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging

Yunyan Zhang [a,b,c,*], Daphne Hong [d], Daniel McClement [c], Olayinka Oladosu [c,e], Glen Pridham [c], Garth Slaney [d]

[a] Radiology, University of Calgary, Alberta, T2N 4N1, Canada
[b] Department of Clinical Neurosciences, University of Calgary, Alberta, T2N 4N1, Canada
[c] Hotchkiss Brain Institute, University of Calgary, Alberta, T2N 4N1, Canada
[d] Departments of Schulich School of Engineering, University of Calgary, Alberta, T2N 4N1, Canada
[e] Department of Neuroscience, University of Calgary, Alberta, T2N 4N1, Canada

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural network
Model interpretation
Magnetic resonance imaging
Heatmap
Class activation mapping/CAM
Gradient-GAM
Grad-CAM++
Disease type
Multiple sclerosis.

ABSTRACT

Background: Deep learning using convolutional neural networks (CNNs) has shown great promise in advancing neuroscience research. However, the ability to interpret the CNNs lags far behind, confounding their clinical translation.

New method: We interrogated 3 heatmap-generating techniques that have increasing generalizability for CNN interpretation: class activation mapping (CAM), gradient (Grad)-CAM, and Grad-CAM++. To investigate the impact of CNNs on heatmap generation, we also examined 6 different models trained to classify brain magnetic resonance imaging into 3 types: relapsing-remitting multiple sclerosis (RRMS), secondary progressive MS (SPMS), and control. Further, we designed novel methods to visualize and quantify the heatmaps to improve interpretability.

Results: Grad-CAM showed the best heatmap localizing ability, and CNNs with a global average pooling layer and pretrained weights had the best classification performance. Based on the best-performing CNN model, called VGG19, the 95th percentile values of Grad-CAM in SPMS were significantly higher than RRMS, indicating greater heterogeneity. Further, voxel-wise analysis of the thresholded Grad-CAM confirmed the difference identified visually between RRMS and SPMS in discriminative brain regions: occipital versus frontal and occipital, or temporal/parietal.

Comparison with existing methods: No study has examined the CAM methods together using clinical images. There is also lack of study on the impact of CNN architecture on heatmap outcomes, and of technologies to quantify heatmap patterns in clinical settings.

Conclusions: Grad-CAM outperforms CAM and Grad-CAM++. Integrating Grad-CAM, novel heatmap quantification approaches, and robust CNN models may be an effective strategy in identifying the most crucial brain areas underlying disease development in MS.

1. Introduction

Deep learning based on convolutional neural networks (CNNs) is playing an increasingly important role in image pattern recognition (LeCun et al., 2015). In neuroimaging, the CNNs have also shown the unprecedented likelihood to advance our disease classification and prediction abilities (Vieira et al., 2017). Multiple sclerosis (MS) is a prevalent inflammatory demyelinating disease that causes both visible and invisible tissue damage in the central nervous system, leading to various types of functional impairment in patients. While most people start with a relatively mild, relapsing remitting form (RRMS), >50% of them worsen to a severe secondary progressive phenotype (SPMS) within 10–15 years of disease onset. While the mechanisms remain unclear, recent evidence suggests that deep learning using a CNN can differentiate MS from controls (Yoo et al., 2018) and distinguish MS subtypes (Marzullo et al., 2019). However, the lack of ability to interpret the CNN models makes it difficult to effectively translate technology into clinical care.

* Corresponding author at: Departments of Radiology and Clinical Neurosciences, University of Calgary, Alberta, T2N 4N1, Canada.
E-mail address: yunyzhan@ucalgary.ca (Y. Zhang).

https://doi.org/10.1016/j.jneumeth.2021.109098
Received 18 August 2020; Received in revised form 4 February 2021; Accepted 7 February 2021
Available online 11 February 2021

A typical CNN contains multiple layers of artificial neurons. Through a step-wise approach, the CNN is capable of detecting increasingly abstract image features automatically (Liu et al., 2019). It is the unique feature patterns learned this way from the input data that permits the CNN to make the best possible decisions. Yet this process also generates enormous numbers of features, making the information extremely compact, particularly in the deep layers when learning progresses. It is nearly impossible for humans to interpret the learning without use of targeted techniques. Over the past few years, several approaches have emerged in this regard. One example is to develop methods showing the texture of imaging features at individual layers to examine the hierarchical process of learning (Zeiler et al., 2011). Alternatively, there have been strong efforts focusing on creating meaningful heatmaps that highlight the importance of individual pixel regions in an input image to its classification using a CNN.

Multiple techniques have shown the feasibility to generate the heatmaps. These include strategies using deconvolution (Zeiler and Fergus, 2014), layer-wise relevance propagation (Samek et al., 2017), and saliency map construction (Ghorbani et al., 2020; Simonyan et al., 2014; Thomas et al., 2020). However, various technical challenges still exist associated with these methods, including the vulnerability to signal noise, lack of sensitivity to input perturbations, and lack of quantitative criteria to assess the quality of backpropagation of information (Ghorbani et al., 2020; Samek et al., 2017; Smilkov et al., 2017). Class activation mapping (CAM) serves as an alternative approach. At its basic form, the CAM requires the addition of an additional pooling layer to the target CNN, and this limits the interpretation to a specific layer only (Zhou et al., 2016). Recently, 2 new variants of CAM have emerged, namely, gradient (Grad)-CAM (Selvaraju et al., 2017), and Grad-CAM++ (Chattopadhyay et al., 2018). Both can interpret arbitrary layers of a CNN, without the need of any architecture modifications, leading to increased flexibility. Moreover, the Grad-CAM++ facilitates higher order gradient computations in heatmap generation, so it may have an even greater generalizability than Grad-CAM. Current literature has already documented the feasibility of using CAM and Grad-CAM to identify classification-relevant CNN features in deep learning (Fernández et al., 2020; Jonas et al., 2019; Rajpurkar et al., 2018). However, there is lack of evidence on the relative ability of the CAM methods, particularly in human studies, and how choices of a CNN impact heatmap outcomes.

Based on clinical brain MRI of MS patients, the purpose of our study was to investigate and compare the utility of the 3 CAM techniques in understanding deep learning with a CNN. In addition, we evaluated 6 CNN models to assess the impact of CNN architecture on heatmap results, and explored the difference in heatmap patterns between patient groups to understand the pathological relevance of our findings based on best combinations of CNN and CAM options, as well as new visualization and quantification approaches of the heatmaps.

2. Materials and methods

2.1. Dataset

This study focused on conventional MRI scans acquired initially for a clinical study aimed at developing quantitative measures of neuroprotection and repair in MS. The dataset included 19 MS patients (10 RRMS, 9 SPMS), and 19 age- and sex-matched controls, all being right-handed females. In addition, to enhance the likelihood of detecting the potential differences in patient outcome between disease forms, the recruitment particularly focused on subjects with distinct demographic characteristics. Compared to SPMS participants, the RRMS participants had a lower disability score (median: 2.0 versus 6.5), shorter disease duration (median: 5 years versus 28 years), and a younger onset age [mean (range): 38.7 (28–53) years versus 58.2 (49–75) years].

The MRI protocol included 3 anatomical sequences typically used in clinical MS imaging: T1-weighted (T1), T2-weighted (T2), and FLAIR MRI. Example imaging parameters were: repetition time (TR)/echo time (TE) = 8/3 ms, 156 slices for T1; TR/TE = 3000/80 ms, 52 slices for T2; and TR/TE = 6000/128 ms, 248 slices for FLAIR. Image standardization involved 4 main pre-processing steps: 1) brain extraction; 2) co-registration, to align T2 and FLAIR to T1 that had the best anatomical contrast; both steps 1 and 2 used the FSL Library (Oxford, UK); 3) image non-uniformity correction, using N4 (Tustison et al., 2010); and 4) signal intensity normalization, to the range 0−1. To improve computing efficiency, we excluded the MRI slices from the very top and bottom areas of the scans that did not show visible brain tissue, so the number of slices reduced from 156 (T1) to 135 per sequence, totaling 135 × 3 images/patient/scan. Further data curation included the following six transformation procedures to augment the samples: rotation range (30°), shear range (0.2), zoom range (0.2); width and height shift (0.1), and horizontal flip.

2.2. CNN classification

This study applied transfer learning using CNNs pre-trained with the large ImageNet dataset (Russakovsky et al., 2014). Transfer learning was beneficial as it could improve model performance based on relatively small sample sizes as seen in the present study, by using knowledge already learned in a similar classification task. Our tests included 6 classification models from a few common CNNs: ResNet50 (He et al., 2015), and the VGG16 and VGG19 variants of the VGG family (Simonyan and Zisserman, 2015). The models differed by the type of initialization: ImageNet weights versus random weights, and type of architecture customization: a global average pooling (GAP) layer versus fully connected (FC) layers used prior to output (Table 1). All programming used the Keras deep learning framework and the TensorFlow backend (Abadi et al., 2016) coded in Python 3.6. Hardware utilized the Tesla K40 m and V100 GPUs built with two cluster computing resources at the local institution for accelerated learning.

We incorporated the data from different MRI sequences as separate channel inputs to a CNN model, similar to the process in handling RGB images. Therefore, each input slice contained a triplet of MR images from T1, T2, and FLAIR respectively, and each model was a 3-class output: RRMS, SPMS, and control. As a standard practice in this field, we divided the MRI data into 3 portions: training, validation, and testing, corresponding to 65%, 15%, and 20%. Data randomization followed a stratified approach to handle the imbalanced data between cohorts, such that each portion contained a similar ratio of images from each class. There was no overlap in patients between portions to avoid duplicate use of images belonging to the same patient.

Model training used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.00002, and categorical cross-entropy as the loss function. The batch size was 16, and the number of epochs was 20 initially. Each model had been through 5 runs under the same procedure (each with a new set of training and validation samples), with each run repeated another 5 times to ensure reliability of the results. In addition, we implemented 'callback' strategies over training/validation as commonly done in the field, which allowed the program to select and save the best model at each repeat, and to proceed to the next repeat even before completing the full 20 epochs if the selection criteria were reached (e.g. early stopping) for maximum efficiency. Finally, to

Table 1
Model performance under different settings.

Model Architecture	ImageNet Weights	Loss	Accuracy (%)
ResNet50 with FC	yes	1.08	51.07
ResNet50 with GAP	yes	1.32	50.12
VGG16 with GAP	**yes**	**0.16**	**93.76**
VGG19 with GAP	**yes**	**0.12**	**95.42**
VGG19 with FC	yes	0.62	74.66
VGG19 with FC	no	0.70	67.84

Note: FC: fully connected; GAP: global average pooling.

simplify the analysis procedure, this study focused on axial MR images alone.

2.3. Heatmap generation

The CAM-based methods had similarities in heatmap generation. But the specific approaches vary as seen below, leading to different flexibility and generalizability.

CAM: This algorithm required a GAP layer, which gave rise to the spatial average of individual feature maps resulting from the last convolutional layer of a CNN (Fig. 1). In a GAP-CNN, the weighted sum of these GAP values determined the final output class of an image slice. Multiplication between the same weighting information generated at image classification, w_k^c, and the corresponding feature maps, A^k, generated the CAM: $L_{CAM}^c = \sum_k w_k^c A^k$. This calculation enabled a direct mapping of the importance of information from output used at decision-making in classification to the individual regions of the image at input.

Grad-CAM: Class weights derived in this method used the equation:

$$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

where **Z** denotes the total number of elements in a feature map. This equation also involved a GAP calculation, but that was independent of the CNN architecture; it used feature maps generated from the gradient values of the classification score, y^c, with respect to the feature maps, A_{ij}^k, in an arbitrarily selected layer. Consequently, heatmap generation used the formula: $L_{Grad-CAM}^c = ReLU\left(\sum_k w_k^c A^k\right)$, where the ReLU (rectified linear unit) function allowed to evaluate features with positive impact only. In this study, we used the last convolutional layer in computing the weights as suggested previously (Selvaraju et al., 2017), where the resulting heatmaps would be most similar semantically to that of CAM for best possible comparisons.

Grad-CAM++: This algorithm derived the weights by equation:

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

The Y^c was a classification score derived by applying an exponential function on y^c, the classification score for Grad-CAM. Compared to y^c, the score, Y^c, would have an increased order of gradients. The weights, w_k^c, were supplemented with up to 3rd order gradients of a classification score with respect to the selected feature maps, thereby supporting increased generalization as compared to Grad-CAM. The heatmap of Grad-CAM++ was: $L_{ij}^c = ReLU\left(\sum_k w_k^c A_{ij}^k\right)$.

Both Grad-CAM and Grad-CAM++ algorithms were compatible with any image classification algorithm, whereas CAM required a GAP layer as part of the CNN architecture. Our pilot experiment showed that Grad-CAM appeared to be better in localizing image regions than Grad-CAM++. Therefore, we applied Grad-CAM to all CNN models under examination, and additionally applied CAM and Grad-CAM++ to the top-2 ranked models for heatmap method comparison.

2.4. Outcome evaluation and statistics

The assessment included 2 parts: the CNN models used for image classification, and the heatmaps generated to interpret the CNNs. For CNN algorithms, we used the validation dataset to assess model generalizability, with the best model determined as one with the lowest categorical cross-entropy loss as suggested previously (Fernández et al., 2020). Assessing performance of the best model used the test dataset based on accuracy and confusion matrices derived for image slices. For the heatmaps, we first evaluated the ability of the CAM methods in highlighting critical brain areas qualitatively for all subject groups, and then quantitatively for the patient cohorts. The latter involved calculation of the 95th percentile threshold of the normalized whole brain

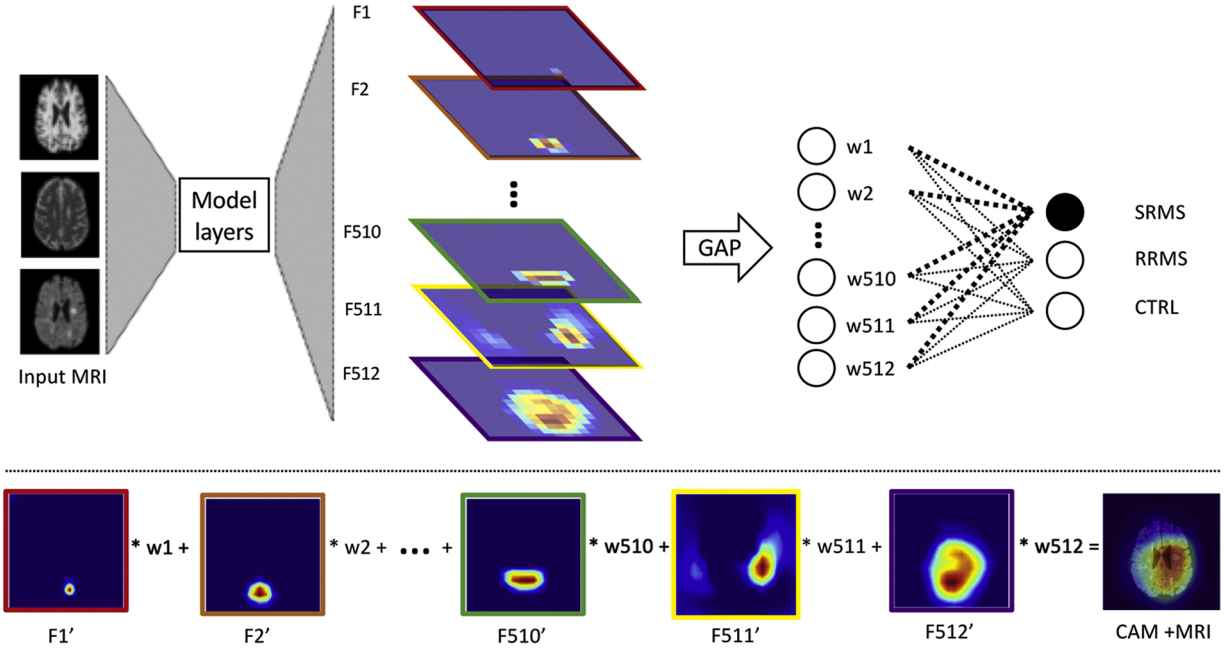

Fig. 1. Schematic diagram of CAM associated with our study based on a VGG model. The input represents 3-channel brain MRI from T1, T2, and FLAIR. Based on the last convolutional layer (block5_conv4), the network generates 512 feature maps (F1-F512) sized 16 × 16 each. The weighted sum of the multiplication between the resized feature maps (F1'-F512', sized 256 × 256 each), and the corresponding weights (w1-w512) used in generating the output class (e.g. SPMS, black dot) produces the CAM image (bottom right). Shown is the CAM overlay with one of the MRI images at input for anatomical correspondence.

heatmaps, and statistical analysis of the threshold values using unpaired Student's *t*-test. Next, using the thresholded heatmaps, we performed voxel-wise statistics as done in fMRI research (Lenoski et al., 2008), which detected the common brain areas highlighted across individuals within a group using one-sample *t*-test, and the different brain areas of highlights between RRMS and SPMS using two-sample *t*-test. All statistical analyses used Python, with $p < 0.05$ as significance, followed by Bonferroni correction for multiple comparisons using adjusted p-values. Finally, following slice-based classification using the top-ranked CNN models, we identified the MRI slices with incorrect predictions in example RRMS and SPMS subjects to further evaluate the validity of the heatmap methods.

3. Results

3.1. MRI characteristics

The original dataset contained 15,390 brain MR images, making 5130 slices of image triplets (3 images per slice from T1, T2, and FLAIR). Of these samples, there were ~9849 images (~3283 slices) for training, ~2462 images (~821 slices) for validation, and 3078 images (1026 slices) for testing, in addition to the multi-fold increased samples through data augmentation. The RRMS and SPMS patients showed variable numbers of focal lesions in brain white matter, best seen in the T2 and FLAIR sequences. There were also different degrees of brain atrophy across subjects, particularly in SPMS, as indicated by the enlarged ventricles and sulci in MRI. There were no visible abnormalities in the MRI of control subjects (Fig. 2). The specific patterns and changes in the images of individual subjects became the basis of feature extraction and reasoning over the deep learning process of the CNNs.

3.2. CNN performance and heatmap quality

The VGG19 model trained with ImageNet weights along with the GAP layer had the best performance, showing the lowest validation loss of 0.12 and highest testing accuracy of 95.42%. Similar settings with the VGG16 model achieved the second best performance, with a validation loss of 0.16 and testing accuracy of 93.76%. These measurements were

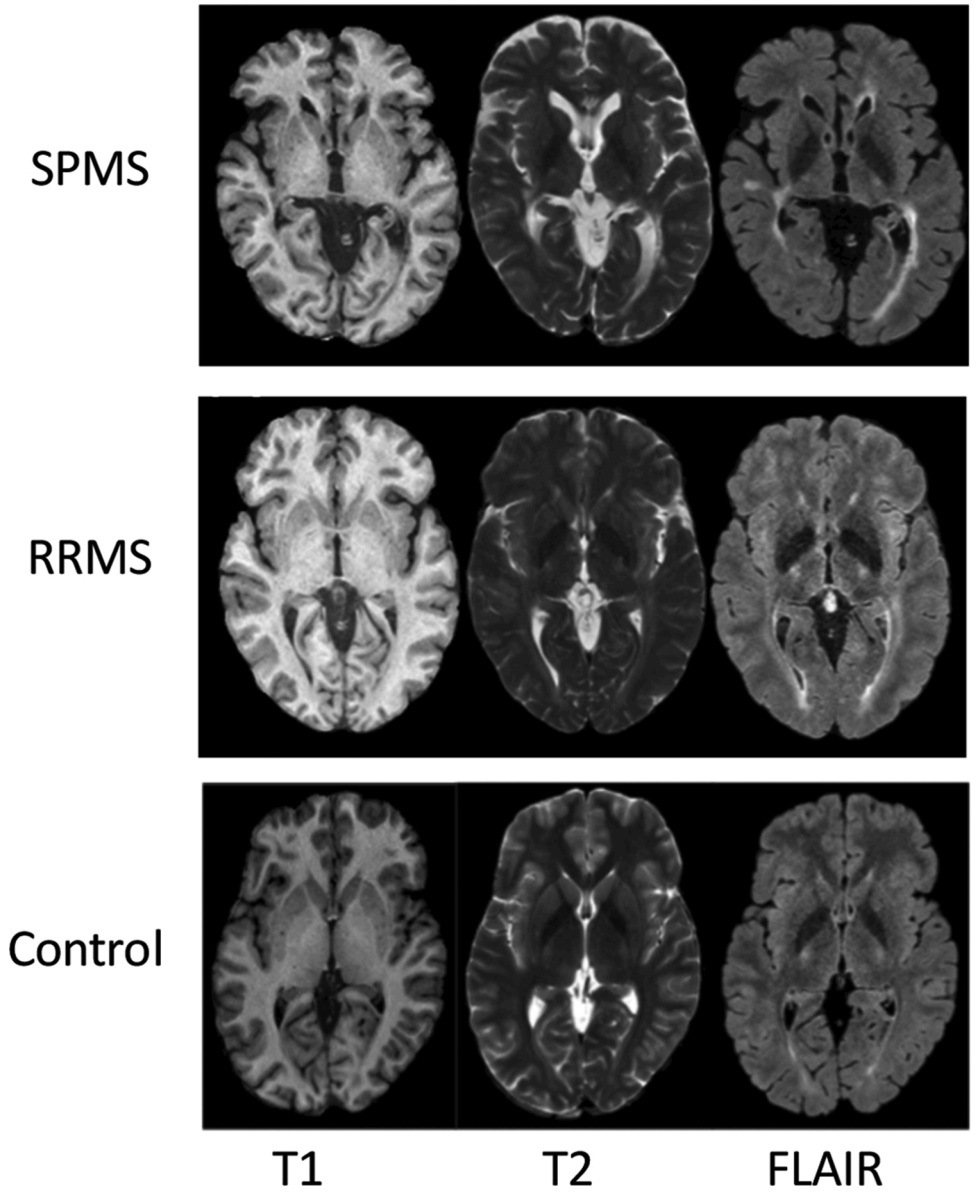

Fig. 2. Example MRI images from each study cohort. Shown are the pre-processed images from T1-weighted, T2-weighted, and FLAIR MRI from a RRMS, SPMS, and control subject, respectively.

better than VGG19 models with customized FC layers, either with or without initialization with ImageNet weights. The two ResNet50 models showed the lowest performance, particularly the one with the GAP layer; both with ImageNet weights initialized (see Table 1). Further evaluation of model performance using confusion matrix demonstrated similar results (Fig. 3).

The quality of the heatmaps associated with individual CNN models also appeared considerably different based on Grad-CAM generated for the same study subject (Fig. 4). In particular, heatmaps of the ResNet50 models (with GAP or FC layers) persistently highlighted the same MRI scan areas regardless of the change in image slices as input, attesting the worst performance. In contrast, the VGG16 model consistently highlighted 2 main brain regions: frontal or occipital lobes, or both, depending on the input of MRI slices. The VGG19 models showed a similar appearance to VGG16, highlighting either occipital, or frontal, or both brain regions. However, VGG19 with GAP showed a higher consistency than VGG16, and the highest consistency among all models across subjects belonging to the same group.

3.3. Differences between CAM, Grad-CAM, and Grad-CAM++ based on top-performing CNNs

The VGG19 and VGG16 CNNs with GAP and ImageNet weights performed the best in image classification, and therefore served as the central means in comparing the 3 heatmap-generating methods. The overall heatmap patterns appeared similar between CAM, Grad-CAM, and Grad-CAM++, but their ability in localizing heatmap regions seemed different (Figs. 5 and 6). The heatmaps generated by Grad-CAM appeared to be the best, with clear highlights of brain areas important to the classification in essentially all subjects. The Grad-CAM++ demonstrated a better background suppressing ability than the CAM in all associated subjects. In addition, while the heatmaps from both appeared similar for SPMS patients, the Grad-CAM++ showed a remarkably better region localizing ability for RRMS subjects than the CAM. Compared to Grad-CAM, there was an obvious generalization phenomenon in Grad-CAM++, which typically highlighted an enlarged territory of the most critical (red) regions. There were no other apparent differences identified between the 3 methods.

3.4. Heatmap differences between subject groups

Based on Grad-CAM of the top-2 VGG classification models, there were considerable differences in the pattern of heatmaps between patient groups. Using VGG16 with GAP, the heatmaps highlighted mostly the frontal or temporal/parietal regions of the brain in SPMS, and highlighted them with similar frequency. In RRMS, the highlights focused on the frontal and occipital regions together or just the occipital region of the brain, at similar frequency. In control subjects, the heatmaps highlighted the middle regions (posterior frontal, parietal, and temporal) with the most frequency.

Using VGG19 with GAP, the patterns were similar to VGG16. In SPMS, the heatmaps highlighted both the frontal and occipital regions together or just the temporal/parietal region the most, and highlighted these 2 patterns with a similar frequency. In both RRMS and control subjects, the heatmap patterns were similar to VGG16 with GAP outcomes. In addition, the cerebellum was also a constant region to be highlighted in all groups but there did not seem to be a constant pattern to distinguish the cohorts.

Quantitative analysis of the heatmaps showed that the mean (standard deviation) 95th percentile value of Grad-CAM in SPMS was significantly higher than RRMS [0.73 (0.06) versus 0.64 (0.03), $p < 0.01$]. In addition, by focusing on subject-wise heatmap values $\geq$ the 95th percentile threshold, voxel-wise analysis detected the most critical brain areas highlighted commonly within a patient group, and areas showing a significant difference when comparing SPMS to RRMS ($p < 0.05$). Subsequent 3D plotting using volume rendering technologies demonstrated that the location of the most significant voxels concentrated in the frontal, parietal/temporal, and occipital areas of the brain (Fig. 7), similar to the visual observations.

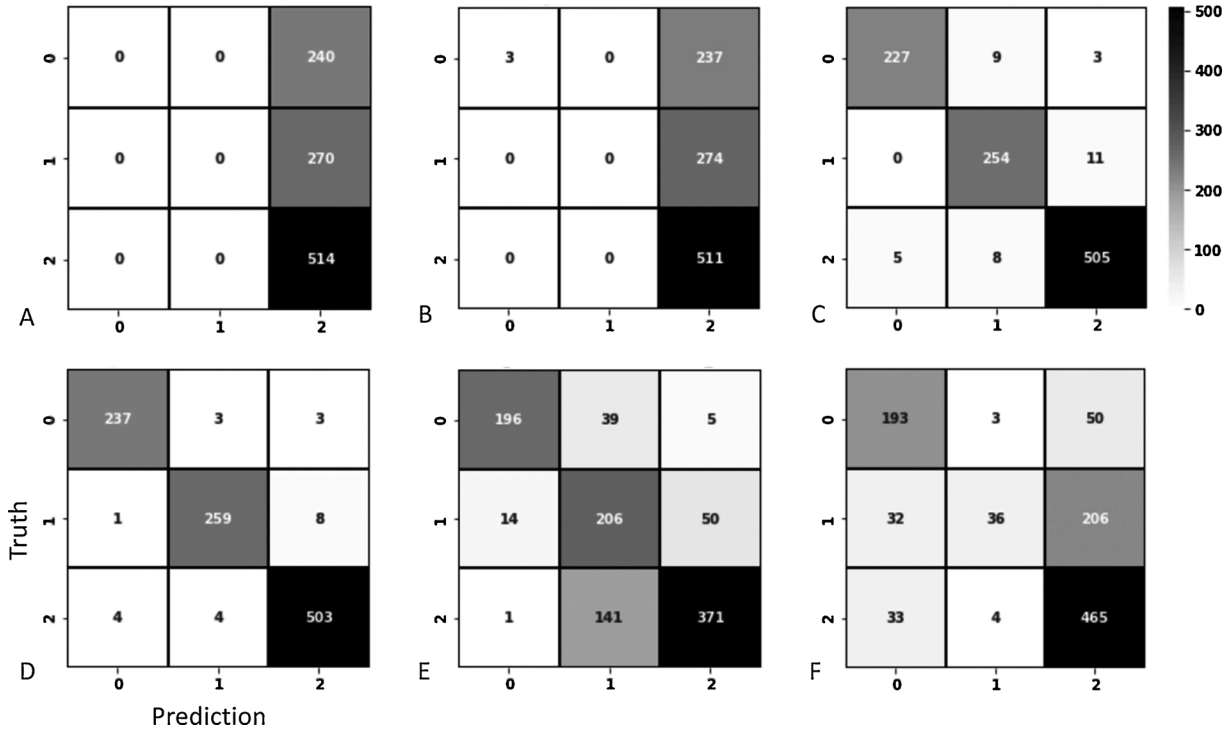

Fig. 3. Confusion matrix results per model. Shown are testing results of each model listed in Table 1, which are ResNet50 with FC (A) and GAP (B) layers, VGG16 and VGG19 with GAP (C-D), and VGG19 with FC layers with (E) and without (F) ImageNet weights. The values in each confusion matrix represent data averaged from 10 tests of the testing data, obtained by splitting the whole dataset 10 times consecutively, each with a new portion of samples but the same sample size.

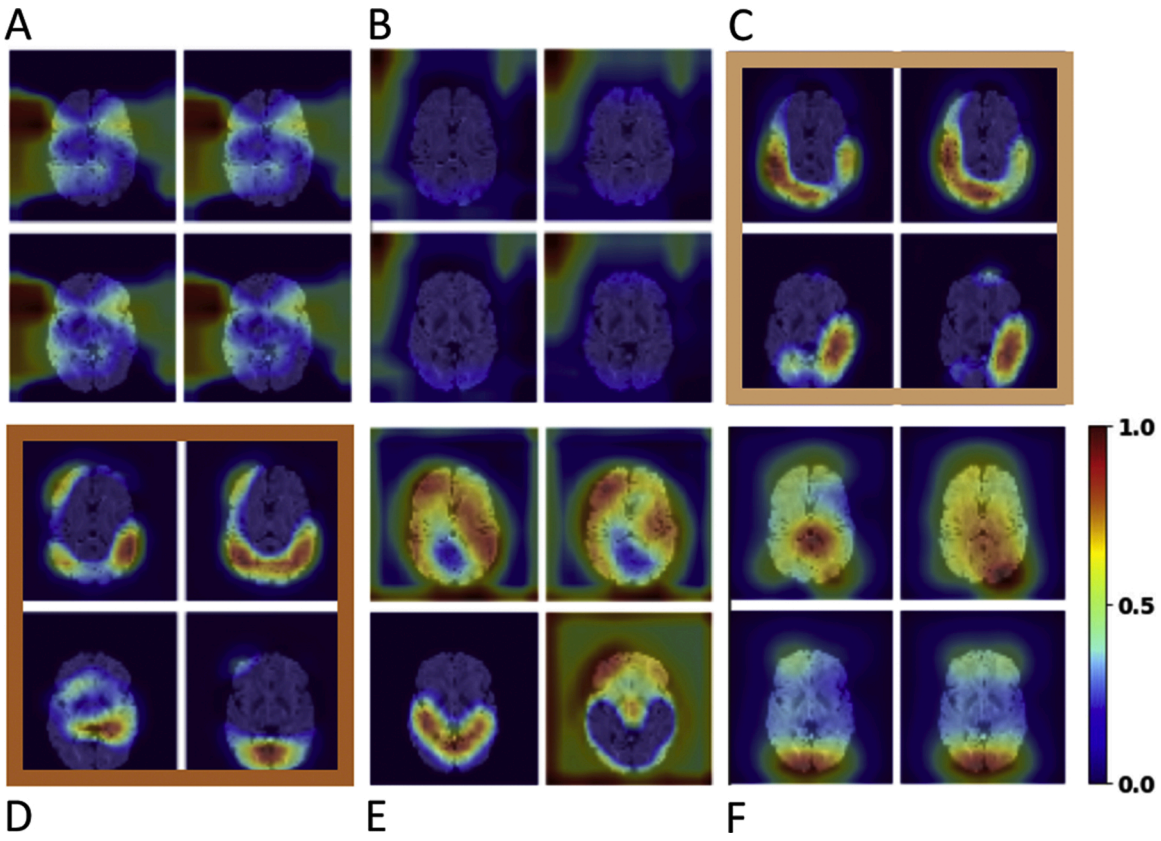

Fig. 4. Example heatmaps. The images are generated using Grad-CAM based on trained classification models for one of the study subjects. The models are: A) ResNet50 with GAP; B) ResNet50 with FC layers; C) VGG16 with GAP; D) VGG19 with GAP; E) VGG19 with FC layers and ImageNet weights; and F) VGG19 with FC layers without ImageNet weights. Color represents the degree of activation from very high (red), high (yellow), to low (green) and very low (blue) respectively. The large light and dark orange boxes highlight the 2 models (C and D) with the best region localization ability for the heatmaps.

3.5. Analysis error interpretation

To further understand the behavior of the CNN models and CAM methods, we also took into account the incorrect class predictions of the images (see Figs. 5 and 6). Based on VGG16 with GAP, MRI slice indices: 45, 46, 48, 49, 89, 99, and 100 from a SPMS patient had high probabilities of belonging to the RRMS class instead of SPMS, which accordingly highlighted the occipital/lower brain regions, supporting the observations for RRMS prediction. The same applied to MRI indices: 65–67, in the VGG19 model with Gap from the same subject. Likewise, in a RRMS subject, MRI slice indices: 63, 64, 82, 83, 84, 88 had high probabilities of being the control class, which highlighted the middle regions of the brain, supporting the observations for control prediction. Further, for the same RRMS subject based on VGG19 with GAP, MRI slice indices: 49, 50, 51, had high probabilities for predicting SPMS, and MRI slice indices: 64 and 84, had high probabilities for predicting control, which highlighted the frontal and middle brain regions accordingly, supporting the observations for SPMS and control subjects. Observation from other subjects showed a similar pattern as seen in these examples.

4. Discussion

Using clinically available brain MRI scans of MS patients, this study evaluated the ability of 3 CAM methods for understanding the mechanisms of deep learning represented by 6 common CNN models for image classification. The results showed that CNN models with a greater classification performance were associated with better heatmap quality, with VGG19 and VGG16 CNNs using a GAP layer and ImageNet weights being the best. Based on these top performing CNN models, Grad-CAM was the best in localizing class-relevant brain regions; Grad-CAM++ seemed to outperform CAM, particularly in localizing RRMS heatmaps. In addition, the heatmap patterns appeared considerably different between subject groups, where the most critical brain areas highlighted in SPMS were much more extensive than RRMS, consistent with the quantitative data.

Interpreting deep learning models has been an ongoing challenge due to the complexity of the learning process. For a classification CNN, the number of parameters increases and the size of feature maps decreases persistently when the network deepens. While highly informative, the parameters themselves as raw data have no direct interpretability; they are just weights in a mathematical function tuned to optimize network decisions. One critical role of the model interpretation methods is to connect the output class with the input image and the features learned to explain the image class, as done by the CAM family.

The CAM-based interpretation techniques are unique as they are entirely model compatible. One major limitation of the vanilla CAM, however, is its requirement to modify the CNN architecture (e.g. the need for a GAP layer), and the subsequent (re)-training of the modified network. The Grad-CAM overcomes these issues and has the flexibility to integrate with any layer of a post hoc network (Selvaraju et al., 2017). Therefore, Grad-CAM is more preferable than CAM, especially given an improved heatmap localization ability. Notably, while being a new polymorphism, Grad-CAM++ seems to be less competitive than Grad-CAM in region localization, likely due to the purposely designed greater generalization potential. The fact that Grad-CAM++ performs relatively better for RRMS than SPMS may suggest that the consequences of generalization impact less on simpler patterns of highlights as seen in RRMS MRI, deserving further validation.

Current results also suggest that the performance of interpretation

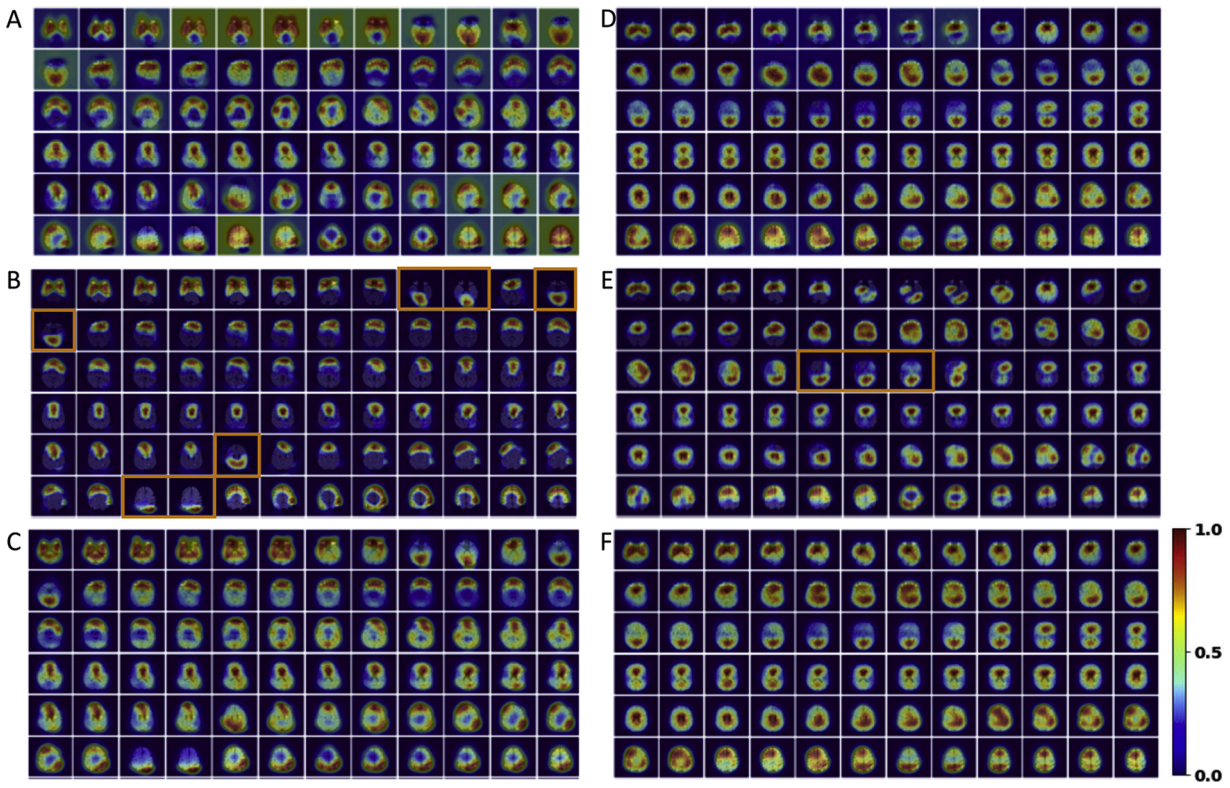

Fig. 5. Heatmaps generated from a SPMS patient. Shown are results from CAM (A, D), Grad-CAM (B, E), and Grad-CAM++ (C, F), obtained using VGG16 with GAP (A-C) and VGG19 with GAP (D-F) respectively. The orange boxes highlight the incorrectly classified MRI slices (to RRMS) along with the corresponding Grad-CAM heatmaps. Each section contains 72 (37 to 108) of the 135 brain MRI slices included in a sequence of the subject, index numbering from top to bottom, and left to right.

methods is related to the type of classification models used. This is in line with expectations from a prior study (Samek et al., 2017), but wherein there was no systemic investigation. In assessing model accuracy of the CNNs, we undertook a slice-by-slice approach. Based on the top-ranked VGG16 and VGG19 models, the correctly predicted image slices showed a similar pattern of highlights in the associated heatmaps in subjects belonging to the same group. In contrast, the most important regions highlighted in the incorrectly classified images were highly consistent with the areas representing the 'wrong' class predicted. Other CNN models with the same setting but decreased accuracy failed to achieve similar performance. Of note, the best trained VGG models with GAP can directly combine with CAM for heatmap generation, or with Grad-CAM to understand the impact of image features from an arbitrary layer on classification. The low performance of the ResNet50 classification models may be due to their remarkably greater architecture depth than the other models tested, causing overfitting with small sample sizes like ours despite the use of transfer learning. In addition, the way of customization may also not be ideal for these deep CNN models in the present study, including the number of nodes in the customized FC layers, and the use of GAP. Finally, except for the classification models and heatmap methods, the quality of input data may also act as a factor impacting heatmap outcomes. In the present study, our data were relatively 'clean' as we used MRI scans acquired from a clinical study using standardized MS imaging protocols, followed by rigorous data curation. While this may not completely equalize signal intensity across images or subjects, other research also suggests that variations in image intensity may not affect deep learning results, but instead may improve model generalizability (Pontalba et al., 2019).

Our overall observations suggest that the heatmap patterns were considerably different between groups. Initially, our sample included patient groups with distinct clinical and demographic characteristics and controls with matched indices. Thus, it is reasonable to expect the detection of structural differences between cohorts. However, it was unclear where the differences would reside in a subject, and by which degree. Integrating heatmap methods into a CNN provides a novel opportunity to interrogate this challenge. Results from our best VGG models along with Grad-CAM found that the territory of the most discriminative brain regions was much more extensive in SPMS, involving nearly the whole brain (frontal, temporal, parietal, and occipital regions), than RRMS that featured mainly the occipital area, and at a lesser extent, the frontal region. Highlight of the temporal/parietal areas in SPMS is unique compared to RRMS. Although the control subjects also show highlights in these 'middle' brain regions, there are no consistent patterns detected in other brain areas in the controls, likely a differentiation from SPMS.

The characteristic patterns in the heatmaps of RRMS and SPMS may indicate an intimate relationship between CNN-extracted features and disease pathology. MS is a complex disease characterized by multiple pathological changes including myelin damage and repair, axonal damage, and inflammation (Correale et al., 2017). The changes may manifest as either focal lesions or lesion-free areas; both can cause various degrees of brain atrophy and functional decline. While there is no direct differentiation of the types of pathology, the most critical brain regions detected in this study overlap more frequently with areas close to sulci and ventricles than lesions, which may indicate a potentially greater role of atrophy. Yet it is worth noting that the observed areas of atrophy can be due to tissue alterations in both local and remote brain regions, deserving further clarification in future studies. Tissue damage in the occipital brain region would lead to vision disturbance, which is one of the most frequently presented symptoms in MS patients, and ~80% of them start with RRMS (Sakai et al., 2011). In this study, our highlight of the occipital region in both qualitative and quantitative analyses of RRMS heatmaps seems to reflect this underlying pathology of the cohort. Similarly, in SPMS, our highlight in frontal,

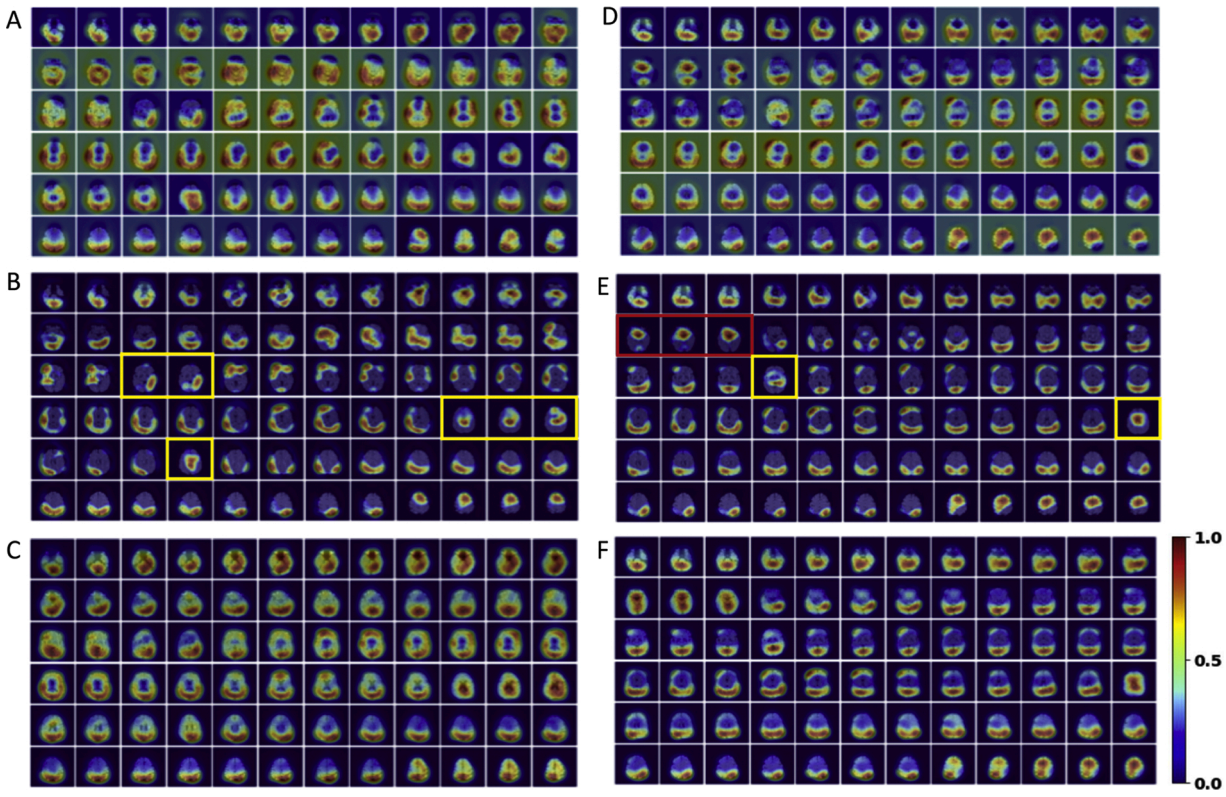

Fig. 6. Heatmaps generated from a RRMS patient. Shown are results from CAM (A, D), Grad-CAM (B, E), and Grad-CAM++ (C, F), obtained using VGG16 with GAP (A-C) and VGG19 with GAP (D-F) respectively. The boxes highlight the incorrectly classified MRI slices (red: SPMS; yellow: control) overlaid with Grad-CAM heatmaps. Each section contains 72 (37 to 108) of the 135 brain MRI slices included in a sequence of the subject, index numbering from top to bottom, and left to right.

temporal/parietal, and occipital regions may indicate tissue injury underlying the 3 key domains of functional deficits in this group: gait, cognition, and vision, respectively (Fox et al., 2012). Further, SPMS is a natural consequence of RRMS. Thus, it is not surprising to observe overlapping changes in tissue pathology, such as highlight of the occipital area in both, and at a lesser extent, the frontal area. However, the SPMS pathology advances with disease development (Bramow et al., 2010), which may be consistent with our quantitative Grad-CAM data showing greater 95th percentile values, suggesting broader areas of abnormality, in SPMS than RRMS. Nonetheless, the mechanisms of disease development remains unclear (Ontaneda et al., 2015). With further confirmation, the areas of difference between RRMS and SPMS identified in this study may become new targets of investigation in this direction, and the image features out of these areas may serve as new user-defined parameters for separate image analysis studies, including classical machine learning for disease characterization.

There are some limitations in this study. The sample size is relatively small. However, our study focuses on the performance of the heatmap generation methods, rather than establishing a new classification model. Moreover, our use of pre-trained CNN models likely helped mitigate the sample size issue, particularly the VGG CNNs. In addition, the CNN models tested are not exhaustive, limiting our scope of data interpretation. Nonetheless, the VGG models are top-ranked in recent large scale pattern recognition competitions, and the CAM and Grad-CAM methods have shown the ability to detect critical discriminating image regions in various studies previously (Fernández et al., 2020; Jonas et al., 2019). Further, there is a notable range of ages associated with individual subjects in the study. While the recruitment has purposely focused on clinical differences between MS groups for best distinction of disease activity, the age range across subjects within a group may have decreased the sensitivity of our analyses, particularly in characterizing subtype-specific effects, given the potential interactions between disease and age related changes in brain structure although it is unavoidable. In the future, we seek to validate our findings using a large sample, including MRI datasets acquired in different planes (e.g. coronal, sagittal) and different parameters, assess the relative importance of pathology types (e.g. lesion versus atrophy) using histology-validated patient images, and test the impact of patient age on the performance of our disease classification and model interpretation methods.

5. Conclusions

Using clinical MRI of MS patients, this study provides proof-of-concept evidence that Grad-CAM is more competitive than Grad-CAM++ and CAM in localizing heatmap patterns, and Grad-CAM++ may be better than CAM. In addition, CNN models with better performance are more capable of generating robust heatmaps, such as the GAP VGG models. Based on Grad-CAM and the best VGG models, the frontal, temporal and parietal brain areas may serve as candidate sites to differentiate RRMS and SPMS, thereby improving our understanding of disease development and the ability of discovering novel imaging biomarkers or treatment targets in MS. Finally, combining 3D volume rendering and quantitative analysis strategies may further promote the utility of Grad-CAM, which in turn should lead to accelerated application of deep learning in clinic settings to advance patient care in both MS and similar diseases.

CRediT authorship contribution statement

Yunyan Zhang: Conceptualization, Formal analysis, Funding acquisition, Resources, Supervision, Writing - original draft, Writing - review & editing. **Daphne Hong:** Formal analysis, Methodology, Software, Visualization, Writing - review & editing. **Daniel McClement:** Data curation, Writing - review & editing. **Olayinka Oladosu:** Data

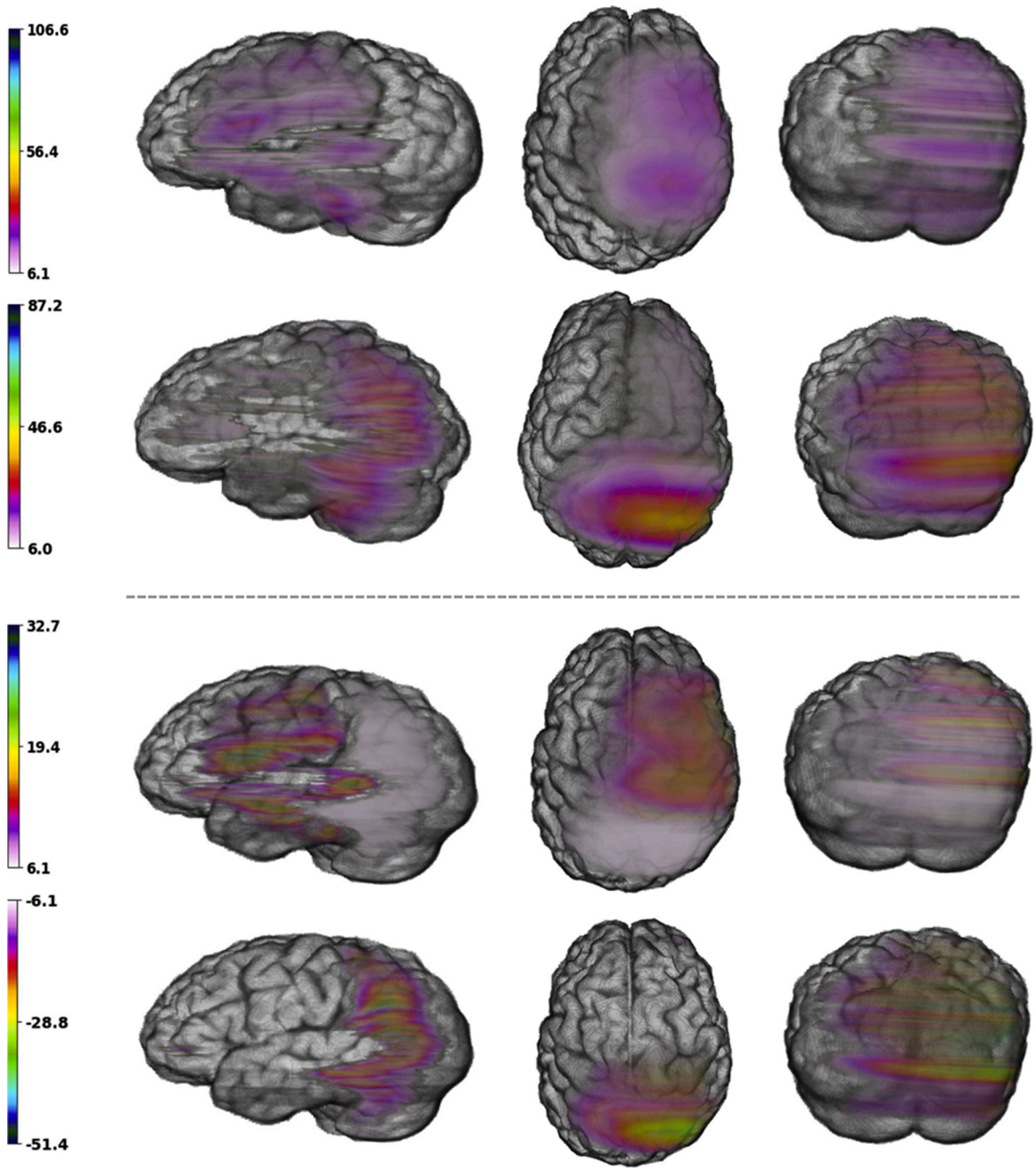

Fig. 7. 3D visualization of Grad-CAM for the RRMS and SPMS groups and group differences. Shown are the T values (plus 95 confidence interval) that demonstrate significance in voxel wise statistics. **Top Panel**: results from one-sample *t*-test, which identify the most critical areas of the brain commonly highlighted across subjects belonging to the same group (top row: SPMS; bottom row: RRMS). **Bottom Panel**: results from two-sample *t*-test, which identify the most critical areas of the brain that distinguish the 2 groups (top row: SPMS > RRMS; bottom row: RRMS > SPMS). The higher the T values, the more significant.

curation, Writing - review & editing. **Glen Pridham:** Data curation, Investigation, Writing - review & editing. **Garth Slaney:** Data curation, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank the patient and control volunteers and the Calgary Clinical MS Research Program (led by Dr LM Metz) for supporting the study, and the intelligent work from Mr Jin Lee associated with this study. We also acknowledge the MS Society of Canada for funding the initial clinical project (PI: Dr LN Brown), and for supporting the work involved in the current study (PI: Dr Y Zhang), along with the Natural Sciences and Engineering Research Council of Canada, andCanadian Institutes of Health Research.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., Google-Brain, 2016. TensorFlow: A System for Large-scale Machine Learning. The

12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16).
Bramow, S., Frischer, J.M., Lassmann, H., Koch-Henriksen, N., Lucchinetti, C.F., Sørensen, P.S., Laursen, H., 2010. Demyelination versus remyelination in progressive multiple sclerosis. Brain 133, 2983–2998.
Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. arXiv.
Correale, J., Gaitán, M.I., Ysrraelit, M.C., Fiol, M.P., 2017. Progressive multiple sclerosis: from pathogenic mechanisms to treatment. Brain 140, 527–546.
Fernández, I.S., Yang, E., Calvachi, P., Amengual-Gual, M., Wu, J.Y., Krueger, D., Northrup, H., Bebin, M.E., Sahin, M., Yu, K.H., Peters, J.M., 2020. Deep learning in rare disease. Detection of tubers in tuberous sclerosis complex. PLoS One 15, e0232376.
Fox, R.J., Thompson, A., Baker, D., Baneke, P., Brown, D., Browne, P., Chandraratna, D., Ciccarelli, O., Coetzee, T., Comi, G., Feinstein, A., Kapoor, R., Lee, K., Salvetti, M., Sharrock, K., Toosy, A., Zaratin, P., Zuidwijk, K., 2012. Setting a research agenda for progressive multiple sclerosis: the International Collaborative on Progressive MS. Mult. Scler. 18, 1534–1540.
Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y., 2020. Deep learning interpretation of echocardiograms. npj Digit. Med. 3, 10.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv.
Jonas, S., Rossetti, A.O., Oddo, M., Jenni, S., Favaro, P., Zubler, F., 2019. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: performance and visualization of discriminative features. Hum. Brain Mapp. 40, 4606–4617.
Kingma, D.P., Ba, J.L., 2015. ADAM: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR 2015).
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lenoski, B., Baxter, L.C., Karam, L.J., Maisog, J., Debbins, J., 2008. On the performance of autocorrelation estimation algorithms for fMRI analysis. IEEE J. Sel. Top. Signal Process. 2, 828–838.
Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., Pietikäinen, M., 2019. From BoW to CNN: two decades of texture representation for texture classification. Int. J. Comput. Vis. 127, 74–109.
Marzullo, A., Kocevar, G., Stamile, C., Durand-Dubief, F., Terracina, G., Calimeri, F., Sappey-Marinier, D., 2019. Classification of multiple sclerosis clinical profiles via graph convolutional neural networks. Front. Neurosci. 13, 594.
Ontaneda, D., Fox, R.J., Chataway, J., 2015. Clinical trials in progressive multiple sclerosis: lessons learned and future perspectives. Lancet Neurol. 14, 208–223.
Pontalba, J.T., Gwynne-Timothy, T., David, E., Jakate, K., Androutsos, D., Khademi, A., 2019. Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks. Front. Bioeng. Biotechnol. 7, 300.
Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., Patel, B.N., Yeom, K.W., Shpanskaya, K., Blankenberg, F.G., Seekins, J., Amrhein, T.J., Mong, D.A., Halabi, S.S., Zucker, E.J., Ng, A.Y., Lungren, M.P., 2018. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 15, e1002686.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Li, F.F., 2014. Imagenet Large Scale Visual Recognition Challenge. arXiv.
Sakai, R.E., Feller, D.J., Galetta, K.M., Galetta, S.L., Balcer, L.J., 2011. Vision in multiple sclerosis: the story, structure-function correlations, and models for neuroprotection. J. Neuroophthalmol. 31, 362–373.
Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K., 2017. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. 28, 2660–2673.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations From Deep Networks Via Gradient-based Localization. arXiv.
Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Proceedings of the International Conference on Learning Representations (ICLR) Workshop, pp. 1–8.
Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR 2015), pp. 1–14.
Smilkov, D., Thorat, N., Kim, B., Viegas, F., Wattenberg, M., 2017. Smoothgrad: Removing Noise by Adding Noise. arXiv.
Thomas, K.A., Kidzinski, L., Halilaj, E., Fleming, S., Venkataraman, G.R., Oei, E.H.G., Gold, G.E., Delp, S.L., 2020. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. Radiology: Artificial Intelligence 2, e190065.
Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.
Vieira, S., Pinaya, W.H., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. 74, 58–75.
Yoo, Y., Tang, L.Y.W., Brosch, T., Li, D.K.B., Kolind, S., Vavasour, I., Rauscher, A., MacKay, A.L., Traboulsee, A., Tam, R.C., 2018. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. Neuroimage Clin. 17, 169–178.
Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. 13th European Conference on Computer Vision, pp. 818–833.
Zeiler, M.D., Taylor, G.W., Fergus, R., 2011. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. 2011 International Conference on Computer Vision, pp. 2018–2025.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929.