applied sciences

MDPI

Article

Multimodal Ensemble-Based Segmentation of White Matter Lesions and Analysis of Their Differential Characteristics across Major Brain Regions

Saima Rathore [1,2,*], Tamim Niazi [3], Muhammad Aksam Iftikhar [4], Ashish Singh [1,2], Batool Rathore [5], Michel Bilello [1,2] and Ahmad Chaddad [3,6,*]

1 Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA 19104, USA; ashish.singh@pennmedicine.upenn.edu (A.S.); michel.bilello@pennmedicine.upenn.edu (M.B.)
2 Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
3 Lady Davis Institute for Medical Research, McGill University, Montreal, QC H3S 1Y9, Canada; tamim.niazi@mcgill.ca
4 Department of Computer Science, COMSATS University Islamabad, Lahore Campus 54000, Pakistan; aksam.iftikhar@gmail.com
5 Department of Psychology, University of Azad Jammu and Kashmir, Muzaffarabad 131000, Pakistan; batool.rathore767@gmail.com
6 School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China
* Correspondence: saima.rathore@pennmedicine.upenn.edu (S.R.); ahmadchaddad@guet.edu.cn (A.C.); Tel.: +1-240-753-9111 (S.R.); +86-150-7730-5314 (A.C.)

Received: 7 February 2020; Accepted: 29 February 2020; Published: 11 March 2020

Abstract: White matter lesions (WML) are common in a variety of brain pathologies, including ischemia affecting blood vessels deeper inside the brain's white matter, and show an abnormal signal in T1-weighted and FLAIR images. The emergence of personalized medicine requires quantification and analysis of differential characteristics of WML across different brain regions. Manual segmentation and analysis of WMLs is laborious and time-consuming; therefore, automated methods providing robust, reproducible, and fast WML segmentation and analysis are highly desirable. In this study, we tackled the segmentation problem as a voxel-based classification problem. We developed an ensemble of different classification models, including six models of support vector machine, trained on handcrafted and transfer learning features, and five models of Residual neural network, trained on varying window sizes. The output of these models was combined through majority-voting. A series of image processing operations was applied to remove false positives in a post-processing step. Moreover, images were mapped to a standard atlas template to quantify the spatial distribution of WMLs, and a radiomic analysis of all the lesions across different brain regions was carried out. The performance of the method on multi-institutional WML Segmentation Challenge dataset (n = 150) comprising T1-weighted and FLAIR images was >90% within data of each institution, multi-institutional data pooled together, and across-institution training–testing. Forty-five percent of lesions were found in the temporal lobe of the brain, and these lesions were easier to segment (95.67%) compared to lesions in other brain regions. Lesions in different brain regions were characterized by their differential characteristics of signal strength, size/shape, heterogeneity, and texture ($p < 0.001$). The proposed multimodal ensemble-based segmentation of WML showed effective performance across all scanners. Further, the radiomic characteristics of WMLs of different brain regions provide an in vivo portrait of phenotypic heterogeneity in WMLs, which points to the need for precision diagnostics and personalized treatment.

Keywords: white matter hyperintensities; segmentation; support vector machines; ResNet; classification

1. Introduction

White matter (WM) hyperintensities or white matter lesions (WMLs) are a common occurrence in a variety of brain pathologies, including infection, issues in the body's immune system, small vessel ischemia, exposure to hazardous chemicals, and more. In many cases, the reason for the development of WMLs is unknown. These lesions show abnormal intensity signals on magnetic resonance imaging (MRI) such as T1-weighted (T1) and fluid-attenuated inversion recovery (FLAIR) MRIs. These WMLs are generally more prevalent in the MRIs of old-age people [1] and accumulating evidence has shown their association with various old-age diseases such as Alzheimer, cognitive deficit, cerebrovascular disease, and other psychiatric disorders [2–8]. Therefore, segmentation and quantification of WMLs is essential to gain an adequate understanding of the relationship between WMLs and old-age disorders.

The quantification of WML load may have diagnostic and prognostic values for individual patients and may lead to personalized medicine for these patients [9]. There is certainly an immediate need to develop computerized methods for segmentation of WMLs in clinical radiology and to develop methods for detecting longitudinal changes in WMLs over the period of time [9,10]. The presence of new lesions in a patient's brain MRI that already has WMLs will prompt the neurologist to consider a change in the regimen of disease-modifying medication. We also note that WML load in specific regions may have both diagnostic and prognostic value in dementia.

The classical segmentation of WMLs is based on manual labeling, which is a time-consuming and laborious process: It has higher inter- and intra-observer variability, and heavily depends on the experience level of the person segmenting the image. Therefore, it is highly advantageous to develop an automatic, reproducible, and faster WML segmentation method. This has been an active area of research, and several automatic segmentation methods have been proposed in the past for delineation of WMLs; some of these automated methods were based on single modality MRI [11,12], whereas others used multimodal MRI [13,14]. Likewise, a subset of these methods has employed voxel-based classification approach [15,16], whereas others have utilized intensity distribution of the FLAIR signal to determine an optimal threshold to segment WMLs [17]. Some recent methods have utilized deep learning approaches, e.g., Pim et al. used multi-scale convolutional neural network (CNN) on T1, FLAIR, and T_1-weighted inversion recovery image as input to segment WMLs and normal brain structures [18], and Rachmadi et al. utilized CNN with global spatial information in MRI with none or mild vascular pathology for delineation of WMLs [19].

An important limitation of existing literature is that ensemble approaches have not been leveraged extensively and segmentation has been carried out either using conventional learning or deep learning approaches. Also, to the best of our knowledge, an analysis of WMLs of different brain regions to provide an insight into the differential characteristics of WMLs growing in different brain regions has not been conducted before.

In this paper, we propose a novel ensemble classification-based segmentation approach [20] for automated segmentation of WMLs (Figure 1). The ensemble combines the strengths of deep learning and classical machine learning by employing base classifiers from each of the categories. Support vector machines (SVM) and residual neural network (ResNet) are used from conventional machine learning and deep learning domains, respectively. Handcrafted and transfer learning-based features are used to train SVM (linear, radial basis function, and sigmoid) models (six models), whereas patches extracted in a certain neighborhood of each voxel are used to train ResNet (five models). We also present a simple post-processing method based on connected-component generation to eliminate false positives. We have evaluated the proposed ensemble model on a publicly available dataset and have shown the generalizability of the model by training and testing the model on data of different institutions. Moreover, we have analyzed the spatial distribution and radiomic characteristics of WMLs in different brain regions in order to find the underlying differences among them.

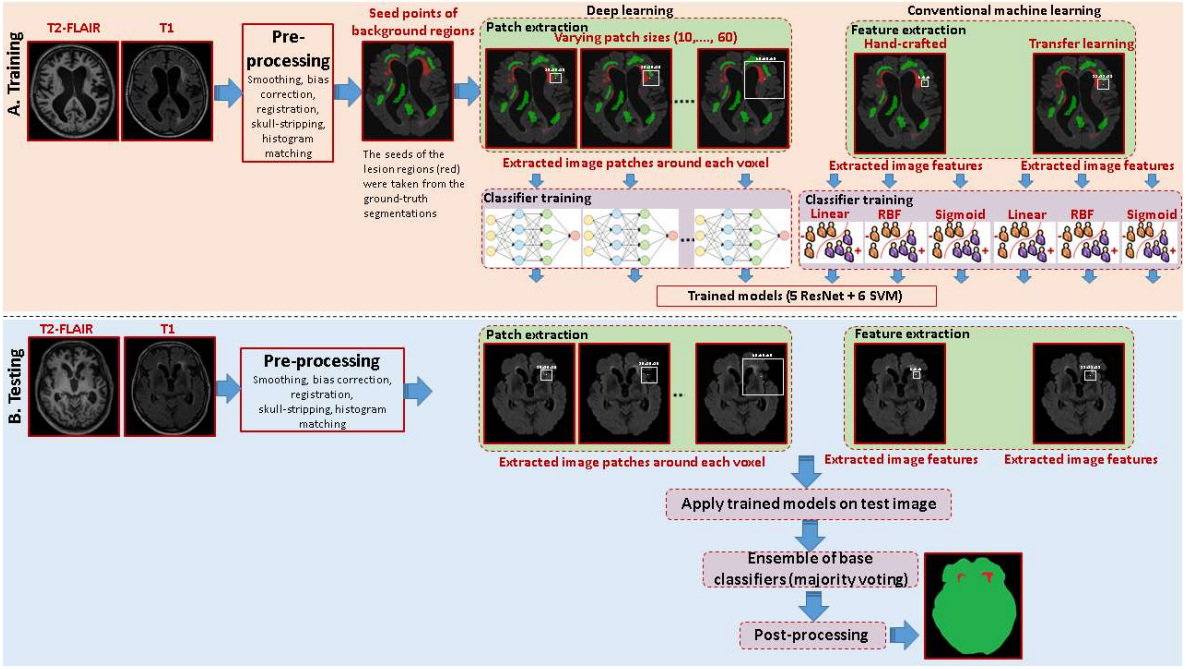

Figure 1. Flowchart of the proposed hybrid model: sequence of steps, starting from the volumetric MRI images (FLAIR and T1) to final segmentation of WMLs. The ground-truth segmented lesions (red) and manually-annotated normal (green) regions were given as input to the segmentation algorithm that uses the features extracted from these regions to train machine learning and deep learning algorithms. The final output of the algorithm is the image segmented into lesion and normal regions.

2. Materials and Methods

2.1. Dataset

The dataset (T1 and FLAIR images) and the corresponding ground-truth were provided by the White Matter Hyperintensity Segmentation Challenge, 2017 [21] (Table 1). The 2017 challenge data comprised 60 training (3 scanners) and 110 testing (5 scanners) images. In this study, we used 60 training and 90 testing images. There were 110 images in the test data; however, we picked only 90 images because they were acquired using the same set of 3 scanners that were used to acquire training data. These datasets are respectively referred as Dataset-I, Dataset-II, and Dataset-III in the rest of the manuscript.

Table 1. Description of the dataset of WMH challenge 2017, as described in the paper summarizing the challenge [21].

Datasets	Site (Institute)	Scanner	Training	Testing
Dataset-I	University Medical Center (UMC) Utrecht, Netherland	3 T Philips Achieva	20	30
Dataset-II	National University Health System (NUHS), Singapore	3 T Siemens TrioTim	20	30
Dataset-III	VU University Medical Centre (VU) Amsterdam, Netherland	3 T GE Signa HDxt	20	30

2.2. Image Preprocessing

The images of each patient were preprocessed using a series of image processing steps, including (i) smoothing (i.e., reducing high frequency noise variations while preserving underlying anatomical structures) using an improved version of non-local means denoising algorithm [22,23], (ii) correction for magnetic field inhomogeneity using N3 bias correction [24], (iii) deformable co-registration of modalities of each patient using Greedy image registration algorithm provided as part of Cancer

Imaging Phenomics Toolkit (CaPTk) [25], (iv) skull stripping using the Multi-Atlas Skull Stripping tool (MUSE) that simultaneously utilizes multiple atlases to strip off the skull region [26], and (v) matching of intensity profiles (histogram matching) of all MRIs of all patients to the corresponding MRIs of a reference patient. Figure S1 in the supplementary material shows MR images (one T1 weighted and one FLAIR image) before and after preprocessing.

2.3. Region Annotation

To train SVM and ResNet classifiers, two classes were required: the lesion class (positive) and non-lesion class (negative). The lesion class voxels were obtained from the ground-truth white matter segmentations, and the non-lesion class voxels were considered to be the left over normal brain voxels. In our case, as the lesion class voxels were significantly smaller in number compared to non-lesion class voxels. Therefore, to provide a balanced training dataset to the classifiers, instead of using all the non-lesion class voxels, the neuro-radiologist annotated some regions in the normal brain. In particular, multiple negative regions (green in Figure 1) were annotated on several tissue types, such as cerebrospinal fluid, gray matter, white matter, etc., to capture the intensity distribution of all the normal tissues.

2.4. Feature Extraction and Classification Using Conventional Models

To feed the data to SVM classifiers, features were extracted from T1, FLAIR, and T1-FLAIR images by using handcrafted and transfer learning approaches. The T1-FLAIR image was obtained after subtracting FLAIR image from the T1-weighted image.

Handcrafted features: An area of $5 \times 5 \times 5$ was selected around each image voxel and 11 features including (i) intensity of images (1 feature); (ii) statistical measures such as mean, median, standard deviation, skewness, and kurtosis (5 features); (iii) gray-level co-occurrence matrix based features, including contrast, correlation, energy, and homogeneity [27] (4 features); and (iv) distance of each voxel from the segmented ventricles (1 feature) were extracted. Ventricles were segmented using Otsu's method, and distance of each voxel from the ventricles was calculated. These features were extracted using several window sizes such as $3 \times 3 \times 3$, $5 \times 5 \times 5$, $7 \times 7 \times 7$, and $9 \times 9 \times 9$, and segmentation accuracy was separately calculated for each window size (supplementary material, Table S1). The segmentation accuracy using window size of $5 \times 5 \times 5$ was higher than that achieved using other window sizes, therefore, we reported results using window size of $5 \times 5 \times 5$.

Transfer learning features: A CNN model that was pre-trained on 1.2 million 3-channel images of the ImageNet LSVRC-2010 (imagenet_vgg_f [28]) was adapted. The model was provided by the VLfeat library [29] as part of their MatLab toolbox (MatConvNet) for computer vision applications. This CNN utilizes multilayer Perceptrons with hidden layers and is a type of deep feed forward neural network. To apply the pre-trained model to our data and extract transfer learning features, a 2-step process was adopted. The first step was to create 3-channel/sliced images using T1, T1-FLAIR, and FLAIR images for each given patient. For instance, for T1, T1-FLAIR, and FLAIR images of size mxnxk = $180 \times 192 \times 256$, k = 256 3-channel images of size mxnx3 = $180 \times 192 \times 3$, where 3 slices/channels of each image were obtained from the corresponding slices in T1, FLAIR, and T1-FLAIR images, were generated. In the second step, window size of $11 \times 11 \times 3$ was considered around each voxel and MatConvNet model was applied, yielding 4096 features for each $11 \times 11 \times 3$ size window, thereby resulting in 4096 features per voxel.

2.5. Regional Patch Extraction and Classification Using ResNet Model

A sliding window mechanism was adopted to extract regional patches of varying size ($10 \times 10 \times 10$,$20 \times 20 \times 20$, … ,$60 \times 60 \times 60$) from MRI, and each patch was assigned the label of its central voxel, i.e., a positive label was assigned to it if the central voxel belonged to WML, and vice versa. A deep ResNet architecture was trained to assign WML and normal label to each patch. ResNet is an advanced artificial neural network, which mimics the working principal of cerebral cortex in the

brain. ResNet utilizes several skip connections in between the adjacent layers. These skip connections between layers add the outputs from previous layers to the outputs of stacked layers. This results in the ability to train much deeper networks than what was previously possible. More detail on ResNet can be found in the pioneering study [30].

We used a modified version of ResNet with architecture of 50 convolutional layers (3×3 kernels, stride of 1, no padding), a max pooling layer with a stride of 2 after every two convolutional layers, followed by a fully connected layer of 512 nodes with a ReLu and a dropout rate of 70%. The network was optimized using stochastic gradient descent optimization method with binary cross-entropy loss. A differential learning rate starting at 5×10^{-5} with cosine annealing was adopted in the beginning; the network kept on varying the learning rate until convergence was achieved or the validation loss began to increase.

2.6. Ensemble Learning

Handcrafted and transfer learning-based features were used to train different kernels of SVM (linear, radial basis function, and sigmoid), thereby leading to six classification models. Similarly, the patches extracted in a certain neighborhood of each voxel were used to train five ResNet models. A majority voting-based ensemble was proposed to combine the outputs of individual SVM and ResNet classifiers.

2.7. Post-processing

A post-processing step was applied at the end to remove false positives. Connected components were generated in the segmented image using 8-neighborhood connectivity, and the components less than 20 voxels in size were subsequently removed. There were few lesions smaller than 20 voxels in size; however, there were many false positives smaller than 20 voxels. Therefore, there was a tradeoff between the model's ability to minimize false positives and to exclude real white matter lesions. To find optimal threshold, the dice score was calculated with an increasing threshold from 5 to 50 voxels. Dice score was gradually improved with increasing threshold and began to drop at 20 voxels. Accordingly, 20 was set as a threshold to remove the false positives.

2.8. Regional Analysis of White Matter Lesions

For analysis of regional characteristics of WMLs, each WML image was mapped to a standard reference template, comprising several delineated brain regions, namely, frontal lobe, temporal lobe, parietal lobe, brain stem, CC fornix, and occipital lobe. To understand the underlying differences between the lesions pertaining to different brain regions, we investigated morphological, intensity, and texture properties of the lesions. We limited our analysis to major brain regions for simplicity. Moreover, frequency distribution of lesions at each voxel of the template was calculated as the ratio of the number of lesions intersecting at that voxel to the total number of subjects.

3. Results

Considering the multi-institutional nature of the dataset, we applied conventional, deep learning, and ensemble models under three different configurations. In the first configuration (scanner-specific configuration), segmentation experiments were conducted using 10-fold cross-validation on data of one scanner at a time. In the second configuration (scanner-agnostic configuration), data from all the scanners was combined pooled/together by ignoring the scanner where the data was coming from, and segmentation training/testing experiments were performed within the pooled dataset using 10-fold cross-validation. In the third configuration (across-scanner configuration), models were trained on data of two scanners at a time and tested on the third scanner. The segmentation experiments in 1st and 2nd configuration were repeated 100 times and average values were reported in the results section.

3.1. Segmentation Performance of Conventional Models

The final segmentation rates obtained using the conventional models are summarized in terms of segmentation accuracy, sensitivity, specificity, Dice score coefficient, and AUC (Table 2, Figure 2). Our model's dice score in scanner-specific configuration on Dataset-I, Dataset-II, and Dataset-III was 80.23%, 81.43%, and 79.12% for linear kernel; 81.34%, 82.65%, and 81.23% for RBF kernel; and 79.23%, 80.43%, and 80.54% for sigmoid kernel, respectively, on handcrafted features. The cross-validated dice score with linear, RBF, and sigmoid kernels was 78.98%, 78.98%, and 77.34%, respectively, when data from different institutions was pooled together (scanner-agnostic configuration).

Table 2. Performance of conventional models on handcrafted and transfer learning features in scanner-specific and scanner-agnostic configurations. I, II, and III correspond to Dataset-I, Dataset-II, and Dataset-III.

Performance Measures	Linear				RBF				Sigmoid			
Handcrafted Features												
	Scanner-Specific			Scanner-Agnostic	Scanner-Specific			Scanner-Agnostic	Scanner-Specific			Scanner-Agnostic
	I	II	III	Combined	I	II	III	Combined	I	II	III	Combined
Dice score	80.23	81.43	79.12	78.98	81.34	82.65	81.23	78.98	79.23	80.43	80.54	77.34
Accuracy	78.21	81.32	79.87	77.45	81.76	81.34	79.23	77.23	77.12	80.12	79.34	76.32
Sensitivity	82.43	78.23	80.12	79.45	78.43	83.32	82.43	80.32	80.54	79.65	81.54	79.12
Specificity	83.12	80.78	81.90	78.43	83.43	80.56	80.56	79.23	77.32	81.87	80.87	76.32
AUC	0.810	0.801	0.805	0.771	0.811	0.807	0.799	0.771	0.721	0.791	0.786	0.789
Transfer learning Features												
	I	II	III	Combined	I	II	III	Combined	I	II	III	Combined
Dice score	82.34	81.76	82.54	78.23	80.54	81.43	80.12	79.76	81.45	82.76	81.67	79.43
Accuracy	83.32	82.65	81.65	79.56	81.56	80.23	80.54	76.23	82.43	81.67	80.43	76.45
Sensitivity	81.42	81.76	80.78	80.23	79.56	79.43	79.67	79.43	80.54	79.32	80.76	79.43
Specificity	81.65	82.87	81.23	77.43	78.43	81.45	81.98	78.12	81.98	81.65	81.78	77.43
AUC	0.823	0.829	0.816	0.812	0.797	0.812	0.814	0.789	0.818	0.823	0.812	0.796

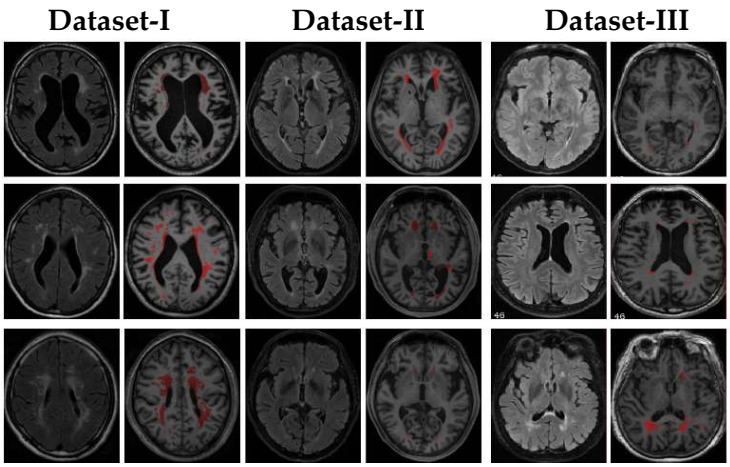

Figure 2. Representative segmentation results of the proposed method. Rows show different subjects, whereas columns show data from different institutions.

Moreover, the model yielded dice score of 82.34%, 81.76%, and 82.54% for linear kernel, 80.54%, 81.43%, and 80.12% for RBF kernel, and 81.45%, 82.76%, and 81.67% for sigmoid kernel, respectively, on Dataset-I, Dataset-II, and Dataset-III by using transfer learning features in scanner-specific configuration. Similarly, the cross-validated dice score using transfer learning features with linear, RBF and sigmoid kernels were 78.23%, 79.76%, and 79.43% in scanner-agnostic configuration. No statistically significant differences of the classification methods among the given datasets by McNemar Test were noted.

3.2. Segmentation Performance of Deep Learning Model

The ResNet model was trained and tested on the patches extracted in the neighborhood of voxels. The cross-validated dice score obtained using ResNet model was 81.93%, 81.65%, and 82.65%, respectively, for Dataset-I, Dataset-II, and Dataset-III in scanner-specific configuration, whereas the dice score of 82.65% was achieved when data from multiple sites was pooled together in scanner-agnostic configuration (Table 3).

Table 3. Performance of ResNet model in terms of various performance measures in scanner specific and scanner-agnostic configurations. I, II, and III correspond to Dataset-I, Dataset-II, and Dataset-III.

Performance Measures	Scanner-Specific			Scanner-Agnostic
	I	II	III	Combined
Dice score	81.93	81.65	82.65	82.65
Accuracy	80.76	82.65	80.76	79.76
Sensitivity	79.23	80.32	83.34	82.34
Specificity	82.87	82.65	81.87	81.87
AUC	0.816	0.815	0.809	0.813

3.3. Segmentation Performance of Ensemble of Conventional and Deep Learning Models

The output of ensemble of SVM models, trained on handcrafted and transfer learning features, and ResNet models was obtained using majority voting (Table 4). The cross-validated accuracy of segmentation was 90.54%, 92.75%, 91.03%, and 92.02%, respectively, on Dataset-I, Dataset-II, Dataset-III, and the combined dataset. Although a reasonably good performance was shown by individual classifiers, higher and statistically significant ($p < 0.01$) accuracy at 95% confidence interval was achieved by the ensemble model. Moreover, comparison of the results of individual models with the multivariate results of combined features and classifiers highlights that the subtle individual features can be synthesized in an index of higher distinctive performance.

Table 4. Performance of the ensemble model in terms of various performance measures in scanner-specific, scanner-agnostic, and across-scanner configurations.

Performance Measures	Scanner-Specific			Scanner-Agnostic	Across-Scanner		
	I	II	III	Combined	I	II	III
Dice score	90.54	92.75	91.03	92.02	90.34	91.36	90.25
Accuracy	89.65	91.98	89.63	91.78	88.76	90.87	89.87
Sensitivity	88.78	93.39	92.89	93.98	92.43	93.69	92.59
Specificity	90.23	90.68	90.87	90.65	87.43	88.36	91.78
AUC	0.916	0.917	0.899	0.907	0.891	0.907	0.799

To verify the generalizability of the proposed ensemble, we evaluated its performance in across-scanner setting, where the model was trained on data of two scanners at a time and tested on the third scanner. A model tested on Dataset-I, Dataset-II, and Dataset-III yielded dice scores of 90.34%, 91.36%, and 90.25%, respectively.

3.4. Analysis of the Effect of Post-processing

The post-processing method was designed to eliminate the connected components (in the segmentation yielded by the previous step) smaller than a certain threshold. To evaluate the effect of post-processing routines, we compared the segmentation performance obtained before and after the post-processing step. The ensemble model performed reasonably well in all the experiments; however, post-processing step further improved the segmentation (Figure 3).

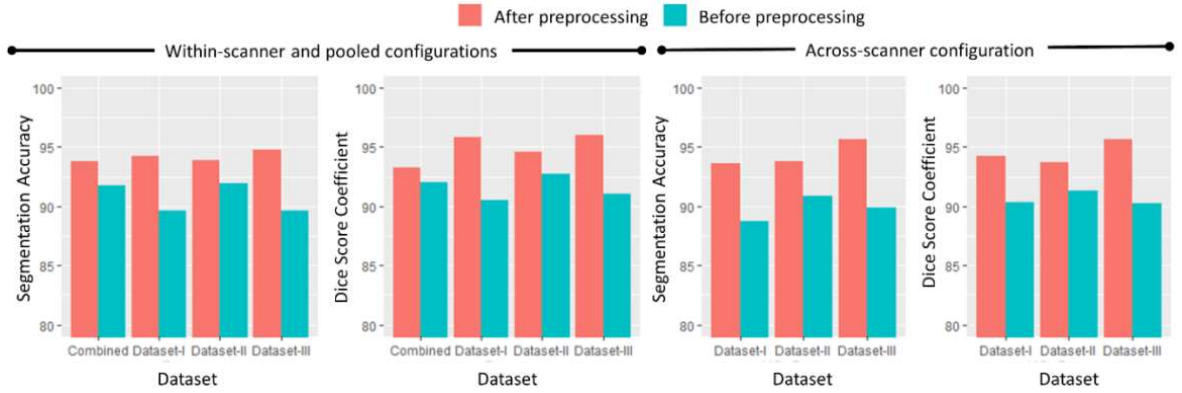

Figure 3. Performance analysis of the proposed method before and after post-processing steps in terms of segmentation accuracy and dice score. The results are provided for the final ensemble model of conventional and deep learning architectures.

3.5. Analysis of the Robustness of Ensemble Model

We also did an assessment of the variation in the predicted labels of each voxel corresponding to all the individual models (Figure 4). The main aim here was to determine the number of cases where the labels predicted by the individual models were matching with the correct label. The y-axis in Figure 4 shows the number of models that provided the same label for a given voxel whereas the x-axis shows correctness count of final label assigned by the ensemble classifier after majority voting. The results show that as the number of models providing the same label for the voxels increase, the correctness of the final label, which is achieved after adding the predictions of all these labels also increases, thereby underscoring the robustness of the proposed ensemble methodology. Only 7 (out of the 11) models are shown here for simplicity.

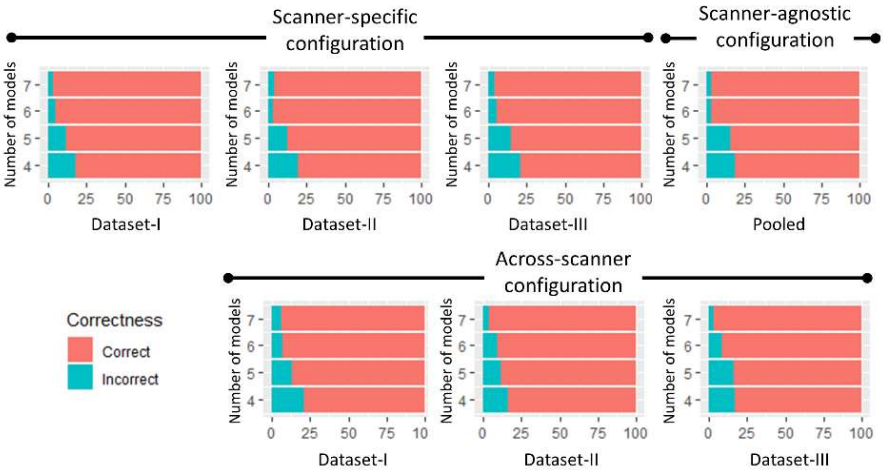

Figure 4. A quantitative evaluation of how many times the individual models provided the same output and whether the consensus of that output was correct or incorrect. The total number of models was 11, however results are shown only for 7 models.

3.6. Analysis of Lesions in Different Brain Lobes

Spatial distribution maps were calculated by mapping the WMLs to a standard atlas template. The maps reveal that most of the lesions are found closer to sub-ventricular zone, i.e., deep within the white matter. The spatial frequency of the lesions goes on decreasing as we move away from the sub-ventricular zone (Figure 5b), and lesions are rarely seen at the edges of white matter. The actual percentage distribution of the lesions in different brain regions (Figure 5a) indicates that most of the

lesions are prevalent in temporal and parietal lobes. A fair number of lesions show predilection to frontal and occipital lobes as well, whereas lesions are less frequently found in fornix and brainstem.

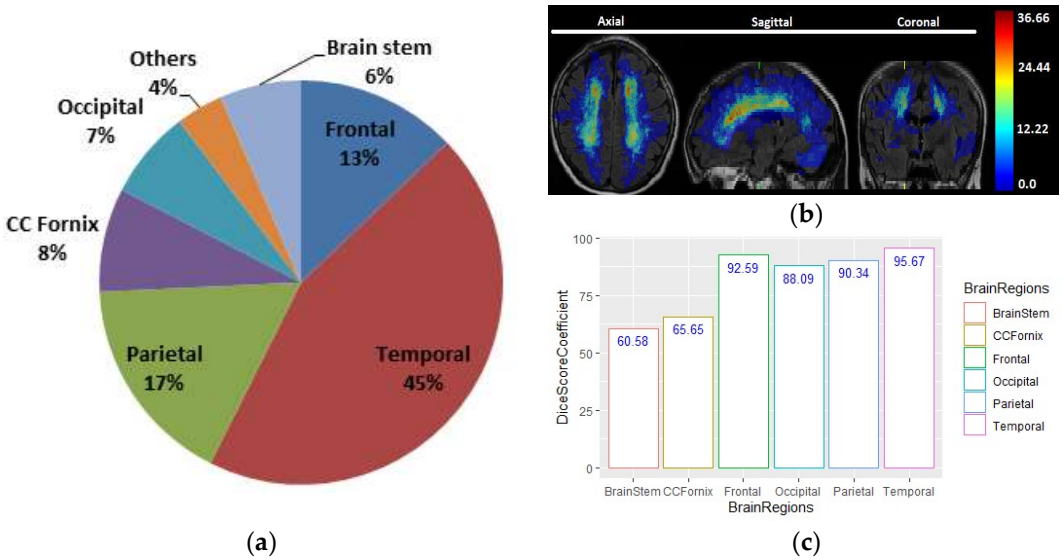

Figure 5. The distribution of lesions in different brain regions. (**a**) Percentage spatial distribution. (**b**) Spatial distribution of lesions mapped to a standard reference template; color bar shows the probability of distribution at each voxel of the reference template. (**c**) Performance analysis of the proposed method in terms of dice score in different brain regions.

An analysis of the segmentation performance in all the brain regions revealed that lesions in temporal and frontal regions of the brain are better segmented compared to lesions of other regions (Figure 5c). The possible reason of the better segmentation performance may be attributed to the overall appearance/shape, signal strength and heterogeneity of the lesions, as reflected by FLAIR and T1 images, in these brain regions.

To reveal the underlying differences in the lesions of different brain regions, we did an analysis of the important radiological characteristics of their lesions. The lesions were supposed to belong to a certain region if their centroid calculated through the morphological operations lied in that region. The extracted radiological measures include average and standard deviation of the FLAIR and T1 intensity, sphericity, normalized volume calculated by the ratio of the number of voxels occupied by a lesion to the total brain size, and texture in FLAIR images as quantified by the measures of entropy and homogeneity (Figure 6). The main characteristics of lesions include the following.

1. **Temporal and parietal lobes:** These lesions showed highest signal strength, irregular shape, largest size, and were moderately heterogeneous.
2. **Frontal lobe:** These lesions were characterized by moderate signal strength, irregular shape, medium size, and moderate heterogeneity.
3. **Occipital lobe:** These lesions had moderate signal strength, moderate irregularity, medium size, and were highly heterogeneous.
4. **CC Fornix and Brain stem:** These lesions were characterized by the least signal strength, high sphericity, smallest size, and had the least heterogeneity.

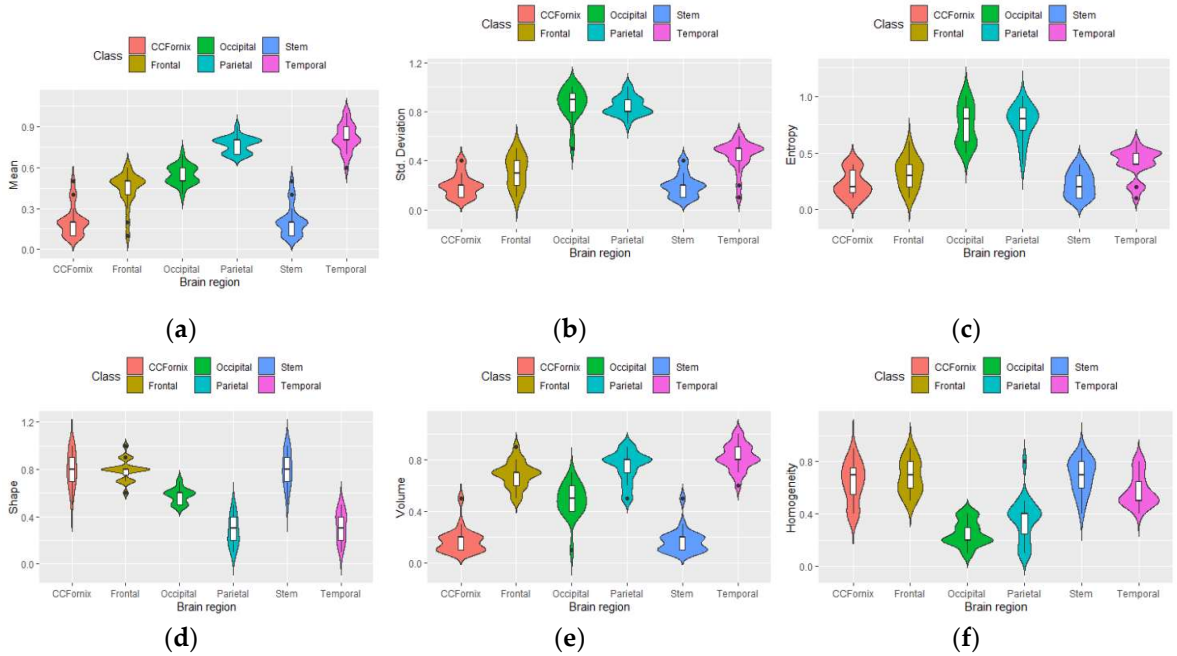

Figure 6. Radiological measures of the lesions of different brain regions. (**a**) Average strength of lesion (FLAIR signal), (**b**) heterogeneity of the lesion (FLAIR), (**c**) entropy/randomness in FLAIR signal, (**d**) shape/appearance of the signal (sphericity), (**e**) size of the lesion, and (**f**) homogeneity in FLAIR signal.

4. Discussion

In this study, we presented a robust WML segmentation method. The final segmentation results are quite encouraging and suggest that the method can be used as an effective tool on an ongoing basis in the clinical routine. The proposed method seems to have a good generalization capability as shown by the testing on completely independent test sets. The ensemble model showed superior performance compared to either of the individual models, thereby highlighting the importance of better results achieved by consensus-based advanced learning approaches. Overall, our methods achieved a high dice score in scanner-specific (Dataset-I = 90.54%, Dataset-II = 92.75%, and Dataset-III = 91.03%), scanner-agnostic (92.02%), and across-scanner configurations (Dataset-I = 90.34%, Dataset-II = 91.36%, and Dataset-III = 90.25%) for WML segmentation. The proposed post-processing step further improved the performance in all the experiments (Figure 3). This highlights that a powerful post-processing step may enhance the segmentation power quite effectively by removing the false positives in such a classification-based segmentation algorithm [31].

It has been shown in the past that radiomic characteristics extracted from MRI may provide important phenotypic/radiographic characteristics of various diseases [32,33]. We also utilized advanced feature extraction and machine/deep learning techniques to comprehensively capture the radiographic characteristics of WMLs using T1 and FLAIR signals. Notably, our approach identified selected radiomic features that are significantly different across lesions of different brain regions, thereby offering noninvasive means of understanding the differences among lesions of different brain regions.

4.1. Importance and Clinical Applicability of the Study

The computational method described in this paper addresses many of the current barriers in clinical radiology. This method offers a standardized approach to resolve intra- and inter-observer variability amongst radiologists in the assessment and quantification of WMLs and is easily implemented even in low resource settings. It is likely that, in the near future, reporting WML load in clinical neuroradiology reports will become standard, at least in certain pathologies such as multiple sclerosis and dementia, as this information has diagnostic and prognostic value [34]. For example, in multiple sclerosis, one could track the load of WML over time to get a temporal assessment of drug efficacy. One could also measure

lesion load in specific white matter structures, e.g., the corpus callosum. The spatial distribution of white matter lesions could have both diagnostic and prognostic value in mild cognitive impairment (MCI) and dementia, for example, WML load in the temporal regions may predict progression to MCI or dementia in asymptomatic patients. Moreover, the amount and size of WML in the periventricular and subcortical regions may predict severity of dementia [35]. As these techniques become more robust, the referring physicians will probably demand this kind of quantification from neuro-radiologists. Importantly, although this study is focused on WML, the same approach could also be used for similar lesion types.

4.2. Analysis of Regional Characteristics

The regional analysis (Figure 6) shows differential characteristics of lesions of different brain regions. According to the current standard practice, all the WMLs are treated the same way, despite their anatomical location and differences in the radiological measures. We believe that the differences in the characteristics of different lesions can provide an insight into the understanding of this disease and may lead towards precision treatment where treatment can be increasingly tailored based on the radiological measures and anatomical location of the lesions. Further, it opens a new direction of research for treatment and care planning of WMLs.

Similar to existing studies in neuro-oncology [33,36,37], the analysis of the radiological measures revealed contrasting differences among the lesions of different brain regions. Most of the WMLs show predilections to temporal part of the brain, especially closer to the subventricular zone, are bright as indicated by the FLAIR signal, moderately heterogeneous, irregular in shape, and heavily diffused and migratory. It is also intriguing to notice that segmentation performance varies across different brain regions; the segmentation accuracy of the lesions in CC fornix and brain stem is lowest, which may be attributed to the lowest signal strength as indicated by the FLAIR signal and the diffused nature of lesions in these regions. The lesions of the temporal and parietal regions are highly bright, whereas the lesions of the CC fornix and brain stem are least bright, thereby showing that the lesions of temporal and parietal regions have more water concentration compared to the lesions of CC fornix and brain stem. Another important finding is the respective size/shape of lesions of different brain regions. For instance, the lesions of temporal and parietal regions, along with being watery, are very diffused, migratory and ill-shaped compared to the lesions of CC fornix and brain stem, which are not only very small in size but also have well-defined boundaries and shape.

4.3. Validation of the Proposed Method across Different Datasets

To attest that our methods would be applicable across multiple institutions, our training and testing datasets were picked from different institutions (across-scanner configuration). The assumption that the combination of features and machine/deep learning classification proposed here allows robust segmentation of WMLs is also supported by the findings in Table 4. Based on the validation of our methods across datasets, our method may perform well in routine clinical settings having much more diverse samples than in controlled experimental settings.

4.4. Limitations and Future Work

One of the limitations of our study is that we used retrospective data; a prospective dataset comparing our methods to standard radiological review would lend further validity to our model. The other limitation is that the sample size is very small. As the strength of deep learning methods often improves with an increase in the number of subjects, having a higher sample size would improve our model. The limited number of patients relative to the number of features utilized in deep learning methods may increase the risk of overfitting. We addressed this potential pitfall by cross-validation of all steps. Prospective validation of our signature on a larger cohort is necessary to establish further reproducibility.

In the future, we aim to develop a deep learning-based method that can tailor the segmentation methodology depending upon the lesions of different brain regions. We also believe that a systematic analysis of the characteristics of lesions of different brain regions via an automatic image analysis framework could lead to a better understanding of the relevant underlying biology, and thus help in precisely and accurately assessing the prediction of the outcomes of lesions.

5. Conclusions

Analysis on MRI has become an important parameter to study diseases with WMLs such as multiple sclerosis. Manual WML segmentation is time-consuming, bears the risk of considerable inter- and intra-rater bias, and may be compromised due to varying WML contrast, particularly when they are subtle. This study presents a new ensemble method for segmentation of WMLs that utilizes the strengths of classical and deep learning paradigms by employing robust base classifiers from each category. The results were fine-tuned in a post-processing step in order to reduce the number of false positives. The proposed rich ensemble was successfully applied to delineate WMLs from FLAIR and T1-weighted images acquired across multiple institutions, thereby emphasizing the generalizability of the proposed ensemble methodology. Lesions in different brain regions showed differential characteristics in terms of signal strength as measured in FLAIR and T1 images, size and shape measures, heterogeneity, and texture, thereby suggesting that lesions in different brain regions arise as a result of different biological process. These results provide a noninvasive portrait of the heterogeneity across the lesions of different brain regions.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-3417/10/6/1903/s1, Figure S1: Preprocessing steps applied on MRI scans, Table S1: Performance of conventional models on hand-crafted features in scanner-specific and scanner-agnostic configurations for varying window sizes.

Author Contributions: Conceptualization, S.R. and A.C.; Data curation, S.R. and M.B.; Formal analysis, S.R., T.N., M.A.I., A.S., A.C., and B.R.; Funding acquisition, T.N.; Methodology, S.R. and A.C.; Project administration, S.R.; Resources, S.R.; Software, S.R.; Supervision, A.C. and M.B.; Validation, S.R., T.N., M.A.I., B.R., A.S., M.B., and A.C.; Writing—original draft, S.R.; Writing—review & editing, S.R., T.N., M.A.I., B.R., A.S., M.B., and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Tamim Niazi.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hopkins, R.O.; Beck, C.J.; Burnett, D.L.; Weaver, L.K.; Victoroff, J.; Bigler, E.D. Prevalence of white matter hyperintensities in a young healthy population. *J. Neuroimaging Off. J. Am. Soc. Neuroimaging* **2006**, *16*, 243–251. [CrossRef] [PubMed]
2. Wen, W.; Sachdev, P.S. Extent and distribution of white matter hyperintensities in stroke patients: The Sydney Stroke Study. *Stroke* **2004**, *35*, 2813–2819. [CrossRef] [PubMed]
3. Fazekas, F.; Chawluk, J.B.; Alavi, A.; Hurtig, H.I.; Zimmerman, R.A. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Ajr. Am. J. Roentgenol.* **1987**, *149*, 351–356. [CrossRef] [PubMed]
4. Dupont, R.M.; Jernigan, T.L.; Butters, N.; Delis, D.; Hesselink, J.R.; Heindel, W.; Gillin, J.C. Subcortical abnormalities detected in bipolar affective disorder using magnetic resonance imaging. Clinical and neuropsychological significance. *Arch. Gen. Psychiatry* **1990**, *47*, 55–59. [CrossRef] [PubMed]
5. Rathore, S.; Habes, M.; Iftikhar, M.A.; Shacklett, A.; Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* **2017**, *155*, 530–548. [CrossRef] [PubMed]
6. Bilello, M.; Doshi, J.; Nabavizadeh, S.A.; Toledo, J.B.; Erus, G.; Xie, S.X.; Trojanowski, J.Q.; Han, X.; Davatzikos, C. Correlating Cognitive Decline with White Matter Lesion and Brain Atrophy Magnetic Resonance Imaging Measurements in Alzheimer's Disease. *J. Alzheimer's Dis. Jad* **2015**, *48*, 987–994. [CrossRef]
7. Chaddad, A.; Desrosiers, C.; Niazi, T. Deep Radiomic Analysis of MRI Related to Alzheimer's Disease. *IEEE Access* **2018**, *6*, 58213–58221. [CrossRef]

8. Chaddad, A.; Toews, M.; Desrosiers, C.; Niazi, T. Deep Radiomic Analysis Based on Modeling Information Flow in Convolutional Neural Networks. *IEEE Access* **2019**, *7*, 97242–97252. [CrossRef]
9. Schmidt, P.; Pongratz, V.; Küster, P.; Meier, D.; Wuerfel, J.; Lukas, C.; Bellenberg, B.; Zipp, F.; Groppa, S.; Sämann, P.G.; et al. Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *Neuroimage Clin.* **2019**, *23*, 101849. [CrossRef]
10. Danelakis, A.; Theoharis, T.; Verganelakis, D.A. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comput. Med Imaging Graph. Off. J. Comput. Med Imaging Soc.* **2018**, *70*, 83–100. [CrossRef]
11. Jack, C.R., Jr.; O'Brien, P.C.; Rettman, D.W.; Shiung, M.M.; Xu, Y.; Muthupillai, R.; Manduca, A.; Avula, R.; Erickson, B.J. FLAIR histogram segmentation for measurement of leukoaraiosis volume. *J. Magn. Reson. Imaging Jmri* **2001**, *14*, 668–676. [CrossRef] [PubMed]
12. Khayati, R.; Vafadust, M.; Towhidkhah, F.; Nabavi, M. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. *Comput. Biol. Med.* **2008**, *38*, 379–390. [CrossRef] [PubMed]
13. Admiraal-Behloul, F.; van den Heuvel, D.M.; Olofsen, H.; van Osch, M.J.; van der Grond, J.; van Buchem, M.A.; Reiber, J.H. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage* **2005**, *28*, 607–617. [CrossRef] [PubMed]
14. Anbeek, P.; Vincken, K.L.; van Osch, M.J.; Bisschops, R.H.; van der Grond, J. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Med. Image Anal.* **2004**, *8*, 205–215. [CrossRef] [PubMed]
15. Maillard, P.; Delcroix, N.; Crivello, F.; Dufouil, C.; Gicquel, S.; Joliot, M.; Tzourio-Mazoyer, N.; Alperovitch, A.; Tzourio, C.; Mazoyer, B. An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology* **2008**, *50*, 31–42. [CrossRef] [PubMed]
16. Herskovits, E.H.; Bryan, R.N.; Yang, F. Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study. *Adv. Med. Sci.* **2008**, *53*, 182–190. [CrossRef] [PubMed]
17. Yoo, B.I.; Lee, J.J.; Han, J.W.; Oh, S.Y.; Lee, E.Y.; MacFall, J.R.; Payne, M.E.; Kim, T.H.; Kim, J.H.; Kim, K.W. Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. *Neuroradiology* **2014**, *56*, 265–281. [CrossRef]
18. Moeskops, P.; de Bresser, J.; Kuijf, H.J.; Mendrik, A.M.; Biessels, G.J.; Pluim, J.P.W.; Isgum, I. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *Neuroimage Clin.* **2018**, *17*, 251–262. [CrossRef]
19. Rachmadi, M.F.; Valdes-Hernandez, M.D.C.; Agan, M.L.F.; Di Perri, C.; Komura, T. Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Comput. Med. Imaging Graph. Off. J. Comput. Med. Imaging Soc.* **2018**, *66*, 28–43. [CrossRef]
20. Rathore, S.; Iftikhar, M.A.; Hussain, M.; Jalil, A. A novel approach for ensemble clustering of colon biopsy images. In Proceedings of the 11th International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 16–18 December 2013; pp. 25–30.
21. Kuijf, H.J.; Casamitjana, A.; Collins, D.L.; Dadar, M.; Georgiou, A.; Ghafoorian, M.; Jin, D.; Khademi, A.; Knight, J.; Li, H.; et al. Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. *IEEE Trans. Med. Imaging* **2019**, *38*, 2556–2568. [CrossRef]
22. Iftikhar, M.A.; Jalil, A.; Rathore, S.; Ali, A.; Hussain, M. Brain MRI denoizing and segmentation based on improved adaptive nonlocal means. *Int. J. Imaging Syst. Technol.* **2013**, *23*, 235–248. [CrossRef]
23. Iftikhar, M.A.; Jalil, A.; Rathore, S.; Hussain, M. Robust brain MRI denoising and segmentation using enhanced non-local means algorithm. *Int. J. Imaging Syst. Technol.* **2014**, *24*, 52–66. [CrossRef]
24. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med Imaging* **2010**, *29*, 1310–1320. [CrossRef] [PubMed]
25. Davatzikos, C.; Rathore, S.; Bakas, S.; Pati, S.; Bergman, M.; Kalarot, R.; Sridharan, P.; Gastounioti, A.; Jahani, N.; Cohen, E.; et al. Cancer Imaging Phenomics Toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. Imaging Spec. Sect. Quant. Imaging Methods Transl. Dev.–Honor. Mem. Dr. Larry Clarke* **2018**, *5*, 011018. [CrossRef] [PubMed]

26. Doshi, J.; Erus, G.; Ou, Y.; Resnick, S.; Gur, R.; Gur, R.; Satterthwaite, T.; Furth, S.; Davatzikos, C. MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage* **2016**, *127*, 186–195. [CrossRef]
27. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *3*, 610–621. [CrossRef]
28. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
29. Vedaldi, A.; Fulkerson, B. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*; ACM: New York, NY, USA, 2010; pp. 1469–1472.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
31. Sousa, A.V.; Mendonca, A.M.; Sá-Miranda, M.C.; Campilho, A. Classification-Based Segmentation of the Region of Interest in Chromatographic Images. In *International Conference Image Analysis and Recognition*; Springer: Berlin, Heidelberg, 2011; pp. 68–78.
32. Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [CrossRef]
33. Rathore, S.; Akbari, H.; Rozycki, M.; Abdullah, K.G.; Nasrallah, M.P.; Binder, Z.A.; Davuluri, R.V.; Lustig, R.A.; Dahmane, N.; Bilello, M.; et al. Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci. Rep.* **2018**, *8*, 5087. [CrossRef]
34. McAleese, K.E.; Walker, L.; Graham, S.; Moya, E.L.J.; Johnson, M.; Erskine, D.; Colloby, S.J.; Dey, M.; Martin-Ruiz, C.; Taylor, J.P.; et al. Parietal white matter lesions in Alzheimer's disease are associated with cortical neurodegenerative pathology, but not with small vessel disease. *Acta Neuropathol.* **2017**, *134*, 459–473. [CrossRef] [PubMed]
35. Targosz-Gajniak, M.; Siuda, J.; Ochudlo, S.; Opala, G. Cerebral white matter lesions in patients with dementia—From MCI to severe Alzheimer's disease. *J. Neurol. Sci.* **2009**, *283*, 79–82. [CrossRef] [PubMed]
36. Zinn, P.O.; Mahajan, B.; Sathyan, P.; Singh, S.K.; Majumder, S.; Jolesz, F.A.; Colen, R.R. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PLoS ONE* **2011**, *6*, e25451. [CrossRef] [PubMed]
37. Bilello, M.; Akbari, H.; Da, X.; Pisapia, J.M.; Mohan, S.; Wolf, R.L.; O'Rourke, D.M.; Martinez-Lage, M.; Davatzikos, C. Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. *Neuroimage Clin.* **2016**, *12*, 34–40. [CrossRef] [PubMed]