Abstract

White matter hyperintensities (WMHs) are foci of abnormal signal intensity in white matter regions seen with magnetic resonance imaging (MRI). WMHs are associated with normal aging and have shown prognostic value in neurological conditions such as traumatic brain injury (TBI). The impracticality of manually quantifying these lesions limits their clinical utility and motivates the utilization of machine learning techniques for automated segmentation workflows. Herein, we develop a concatenated random forest framework with image features for segmenting WMHs in a TBI cohort. The framework is provided publicly through the Advanced Normalization Tools (ANTs) and ANTsR toolkits. MR (3D FLAIR, T2-, and T1-weighted) images from 24 service members and veterans scanned in the Chronic Effects of Neurotrauma Consortium's (CENC) observational study were acquired. Manual annotations were employed for both training and evaluation using a leave-one-out strategy. Lesion load and overlap evaluative comparisons are complimented by feature rankings which showcase the utility of the concatenated approach. Our findings suggest supervised learning methods may be applied to quantify WMHs on routine brain imaging. Paired with correlative outcome data, supervised learning methods may allow for identification of imaging features predictive of diagnosis and prognosis in individual TBI patients.

Introduction

White matter hyperintensities in TBI

White matter hyperintensities (WMHs) are foci of abnormally increased signal intensity seen within white matter regions within the cerebrum and brainstem on fluid attenuation inversion recovery (FLAIR) magnetic resonance imaging (MRI) sequences. WMHs are a frequent finding following traumatic brain injury (TBI) and have been correlated with functional outcome and injury severity in both pediatric[1,2] and adult[3–6] populations.

Further research involving WMHs has shown that regional distribution and volume of WMHs have been shown to possess prognostic value in the TBI patient[2,6–8]. Specifically, lesion volume in corpus callosum correlates with functional scores in the acute phase following injury, while lesion volume in frontal lobes correlates with scores at 1 year following injury[6]. Further, volume of FLAIR lesions within the corpus callosum, brainstem, and thalamus in patients with severe TBI correlates with Glasgow Outcome-Extended (GOS-E) scores—a numeric groupwise assessment used to classify "outcome" in TBI patients where "outcome" refers to the spectrum of possible prognoses from death to disability to recovery.[4]. Additionally, in patients who are comatose following severe TBI the regional distribution of FLAIR lesions within the pons, midbrain, hypothalamus, basal forebrain, parietal, temporal, occipital lobes, and insula along with the observation of grasping or chewing behavior are associated with poor outcome[7].

Despite the above findings which demonstrate that WMHs have potential prognostic value, they are not routinely employed as a diagnostic measure in clinical practice. Performing a comprehensive manual counting of number and distribution of lesions in the clinical setting is simply not practical. As such, the development of automated methods for the rapid identification and quantification of WMHs within individual patients may allow for identification of correlative patterns between WMH number, volume, distribution, and disease state. Further, the development of such lesion quantification approaches may allow for the practical inclusion of this type of information within routine radiological practice. In this work, we present an automated framework for

quantification of WMHs in multi-modal MRI using the random forest machine learning technique.

Random forests for WMH segmentation

The random forests framework[9] is a popular machine learning technique that has demonstrated significant utility for supervised segmentation tasks (e.g., normal human brain segmentation[10]) and other computer vision applications (e.g.,[11]). Random forest-based paradigms have been successfully employed in the delineation of other neuropathologies[12–17] for both single and multi-modal acquisition protocols.

Random forests are conceptually straightforward[9]. They consist of ensembles of decision trees that are built from training data. Once the ensemble of decision trees is constructed, data to be classified is "pushed" through each decision tree resulting in a single classification "vote" per tree. These votes are then used for regression or classification of the data. Although decision trees had been extensively studied, the success of employing collections of such weak learners for boosting machine learning performance (e.g., AdaBoost[18,19]) influenced the similarly sytled conglomeration of decision trees into "forests" with randomized node optimization[20,21]. Finally, Breiman[9] improved accuracy by random sampling of training data (i.e., "bagging") resulting in the current random forest technique applied here.

In this work, we develop a concatenated random forest framework with a feature image set (both spatial and intensity-based) for segmenting WMHs in a large TBI cohort. The entire framework is built on the well-known open-source Advanced Normalization Tools (ANTs)[1] and ANTsR[2] toolkits. Further motivating this research is the availability of several large publicly available imaging data sets that permits testing reproducibility of this automated routine for WMH segmentation and quantification.

[1] https://github.com/stnava/ANTs

[2] https://github.com/stnava/ANTsR

Materials and Methods

Imaging

MR images utilized for this initial report were acquired from a single scanner involved in the Chronic Effects of Neurotrauma Consortium's (CENC) observational study (see Walker et al., this issue). Briefly, participants were Operation Iraqi Freedom/Operation Enduring Freedom (OIF/OEF) era Service Members and Veterans between the ages of 18-60 years with prior combat exposure and deployment(s). The feature images were derived from MR acquisitions of 26 subjects aged 39.6 ± 8.1 years (range 28–58 years). Within this cohort, 24 (92%) were considered positive for TBI based upon the potential concussive events (PCE) interview process described in detail in Walker et al., this issue). Each of the participants that were selected from the larger cohort of participants in this study demonstrated at least one white matter hyperintensity (but as many as 20) on FLAIR imaging.

Images were acquired on a Philips 3.0T Ingenia system with an 8-channel SENSE head coil (Philips Medical Systems, Best, Netherlands). 3D FLAIR sequences were acquired with a turbo spin echo inversion recovery sequence with the following parameters: repetition time (TR) = 4800 ms, echo time (TE) = 325 ms, inversion time (TI) = 1650 ms; 170 sagittal slices with a 1.2 mm slice thickness, 256 × 256 acquisition matrix, and 256 × 256 mm FOV. 3D T1-weighted sequences were acquired with a fast field echo (FFE) sequence with the following parameters: TR = 6.8 ms, TE = 3.2 ms, echo train length (ETL) = 240; Flip angle = 9°, 170 sagittal slices with a 1.2 mm slice thickness, 256x240 acquisition matrix, and 256 × 256 mm FOV. In addition, 3D T2-weighted images were acquired with a turbo spin echo sequence with the following parameters: TR = 2500 ms, TE = 245 ms, ET: = 133; 170 sagittal slices with a 1.2 mm slice thickness, 256 × 256 acquisition matrix, and 256 x 256 mm FOV.

The first author (J. R. S.) performed the manual WMH tracings for all 26 subjects. J. R. S. is a radiologist certified by the American Board of Radiology, with a certificate of advanced qualification in vascular and in-

terventional radiology, over 18 years of research experience in TBI, and 6 years of clinical imaging experience. All multi-modal MR dicom image slices were converted to the nifti file format.[3] Given the utility of the FLAIR sequence in detecting white matter lesions in TBI[3], all nifti image volumes for each subject were rigidly aligned to the FLAIR image of that subject using the ANTs software[22] . The normalized MRI volumes were then provided to J. R. S. who traced each lesion using the ITK-SNAP tool[23] which has multi-image overlay capabilities for visualizing all modalities in all three canonical views.

Quantitative analysis

Figure X provides a graphical overview of the proposed workflow. The major components include offline generation of symmetric multimodal templates, the creation of feature images from the training data which are then used to model the statistical prediction using a concatenated random forest framework. Once these offline steps are performed, a new, unsegmented subject can then be processed using the proposed pipeline.

Feature images for WMH segmentation

Crucial to these supervised segmentation approaches are the creation and selection of "features" as input (i.e., feature images constructed from the training data) in conjunction with expertly identified structures of interest (i.e., WMHs) for model construction. For the targeted application in this work, tissue classification is performed at the voxelwise level. In other words, each voxel within the region of interest is sent through the ensemble of decision trees and receives a set of classification votes from each tree thus permitting a regression or classification solution. Since this procedure is performed at the voxelwise level, intensity information alone is insufficient for good segmentation performance due to the lack of spatial context. For example, as pointed out in[24], higher intensities can be found at the periventricular caps in normal subjects which often confounds automated lesion detection algorithms. Other potential confounds include MR signal inhomogeneity and noise. Therefore, even though machine learning and pattern recognition techniques are extremely powerful and have significant potential, just as crucial to outcome is the creative construction and deployment of salient feature

[3] http://nifti.nimh.nih.gov/nifti-1

images which we detail below.

Supervised methodologies are uniquely characterized, in part, by the feature images that are used to identify the regions of interest. In Table 1, we provide a list and basic categorization of the feature images used for the initial (i.e., Stage 1—more on the use of multiple random forest stages below) segmentation of the WMHs. In addition Figure 1 provides a representation of a set of feature images for a single subject analyzed in this work. Note that in this work we categorize the brain parenchyma with seven labels:

- cerebrospinal fluid (label 1),
- gray matter (label 2),
- white matter (label 3),
- deep gray matter (label 4),
- brain stem (label 5),
- cerebellum (label 6), and
- white matter hyperintensities (label 7).

As mentioned previously, input for each subject comprises FLAIR, T1-, and T2-weighted acquisitions. The T1 and T2 images are rigidly registered to the FLAIR image using the open-source Advanced Normalization Tools (ANTs)[22]. The aligned images are then preprocessed using the denoising algorithm of[25] followed by N4 bias correction[26] which are then normalized to the intensity range $[0, 1]$. Although we could have used an alternative intensity standardization algorithm (e.g.,[27]), we found that a simple linear rescaling produced better results similar to previous work[17].

The T1 image is then processed via the ANTs brain extraction and normal tissue segmentation pipelines[28]. The result is a mask delineating the brain parenchyma and probabilistic estimates of the CSF, gray matter, white matter, deep gray matter, brain stem, and cerebellum[29]. These provide the expertly annotated labels for the first six tissue labels given above. Tissue prior probability maps for segmentation are from multi-model optimal symmetric shape/intensity templates[17] created from the public MMRR data set[30] (cf Figure 3).

Feature values include the preprocessed FLAIR, T1, and T2 image voxel intensities. We also calculate a set of neighborhood statistics (mean, standard deviation, and skewness) feature images using a Manhattan radius of one voxel given the typical size of individual WMHs. For each of the preprocessed images, we calculate the difference in intensities with the corresponding warped template component. Previous success in the international brain tumor segmentation competition[31] was based on an important set of intensity features that were created from multi-modal templates mentioned previously[17] and listed in Table 1. We employ the same strategy here.

To take advantage of the gross bilateral symmetry of the normal brain (in terms of both shape and intensity), and the fact that WMHs do not generally manifest symmetrically across hemispheres, we use the symmetric templates to compute the contralateral intensity differences as an additional intensity feature.

The segmentation probability images described above are used as feature images to provide a spatial context for the random forest model prediction step. Additional spatial contextual feature images include the distance maps[32] based on the csf, gray matter, and deep gray matter images. These latter images are intended to help distinguish white matter hyperintensities from false positives induced by the partial voluming at the gray/white matter interface. A third set of images are based on the voxel location within the space of the template. Similar feature images were used in[33] although, unlike the proposed framework, this previous work lacks normalization to the standard coordinate system provided by the template to dramatically improve spatial specificity across all subjects. To generate these images, the T1 image of each subject is registered to the T1 template component using a B-spline variant[34] of the well-known ANTs Symmetric Normalization (SyN) algorithm[35]. Using the derived transforms, the template coordinate images are warped back to the space of the individual subject.

Stacked (concatenated) random forests for improved segmentation performance

In previous brain tumor segmentation work[17], it was demonstrated that a concatenated supervised approach, whereby the prediction output from the first random forest model serves as partial input for a second random

forest model, can significantly improve segmentation performance. We do the same thing for the work described here where we employ two stacked random forests (or two "stages"). The Stage 1 feature images of the training data (as described previously) are used to construct the Stage 1 model. The training data Stage 1 features are then used to produce the voxelwise "voting maps" (i.e., the classification count of each decision tree for each tissue label) via the Stage 1 random forest model. All the Stage 1 features plus the Stage 1 voting maps are used as input to the Stage 2 model. In addition, we use the Stage 1 voting maps as tissue priors (i.e., probabilistic estimates of the tissue spatial locations) for a second application of the 6-tissue segmentation algorithm with an additional Markov Random Field spatial prior (MAP-MRF)[29]. In order to maximize the spatial information for the n-tissue segmentation process following the voxelwise RF classification of Stage 1, we use all three aligned preprocessed images for multivariate segmentation during the second stage. The resulting seven posterior probability images constitute a third additional feature image set for Stage 2.

Implementation

As pointed out in a recent comprehensive lesion segmentation review[36], although the number of algorithms reported in the literature is quite extensive, there were only four publicly available segmentation algorithms at the time of writing this article. In contrast to the current work, none are based on supervised learning. As we did for our brain tumor segmentation algorithm[17], all of the code described in this work is publicly available through the open-source ANTs/ANTsR toolkits. Through ANTsR (an add-on toolkit which, in part, bridges ANTs and the R statistical project) we use the *randomForest* package[37] using the default settings with 2000 trees per model and 500 randomly selected samples per label per image. Note that we saw little variation in performance when these parameters were changed (i.e. up to 1000 random samples and as little as 1000 trees) which is consistent with our previous experience.

In addition, similar to our previous offering,[4] we plan on creating a self-encapsulated example to showcase

[4] https://github.com/ntustison/ANTsAndArboles

the proposed methodology which will also be available on github.[5] The fact that the data will also be made available through the Federal Interagency Traumatic Brain Injury Research (FITBIR) repository along with the manual labelings will facilitate reproducibility on the part of the reader as well as any interest in extending the proposed framework to other data sets.

Evaluation protocol overview

In order to evaluate the protocol described, we performed a leave-one-out evaluation using the data acquired from the 24 subjects described above. Initial processing included the creation of all Stage 1 feature images for all subjects. The initial brain segmentation of each T1 image and the manual white matter hyperintensity tracings were combined to provide the truth labels for the training data. The "truth" labels are the seven anatomical regions given above.

The leave-one-out procedure is as follows:

- Create Stage 1 feature images for all 24 subjects.
- For each of the 24 subjects:
  - sequester the current subject and corresponding feature images.
  - construct the Stage 1 random forest model from the remaining 23 subjects.
  - apply the Stage 1 random forest model to the feature images of the 23 training subjects.
  - the previous step produces the Stage 1 voting maps for all seven labels.
  - for each of the 23 subjects, perform a Bayesian-based segmentation with an MRF spatial prior using the seven voting maps as additional tissue priors.
  - construct the Stage 2 random forest model from all the Stage 1 feature images, seven voting maps, and seven posterior probability maps from the previous step.
  - send the sequestered subject through the random forest models for both stages.
  - compare the final results with the manually-defined white matter hyperintensity regions.

[5] https://github.com/ntustison/WatchMeHyperventilate

Results

Ranking feature importance

After performing the leave-one-out evaluation, we calculated the *MeanDecreaseAccuracy* feature values for each of the 24 subjects $\times$ 2 models per subject $= 48$ total models. This measure (per feature, per model) is calculated during the out-of-bag phase of the random forest model construction and quantifies the decrease in prediction accuracy from omitting the specified feature. In other words, this quantity helps determine the importance of a particular feature and, although we save such efforts for future work, this information provides us with guidance for future feature pruning and/or additions.

The resulting rankings for both Stages are given in Figures 4 and 5 where the values for the separate stages are averaged over the entire corresponding model set. In addition, we track the variance for each feature over all models to illustrate the stability of the chosen features during the evaluation. This latter information is illustrated as horizontal errors bars providing the 95^{th} percentile Note that the reader can cross reference Table 1 for identifying corresponding feature types and names.

Additionally, it is interesting to note some of the other top performing features for Stage 1. The contralateral difference FLAIR image is highly discriminative over the set of evaluation random forest models (see Figure 6). This accords with the known clinical relevance of FLAIR images for identifying white matter hyperintensities and the fact that such pathology does not typically manifest symmetrically in both hemispheres. Interestingly, the posterior maps for the deep gray matter are extremely important for accurate white matter hyperintensity segmentation. Perhaps the spatial specification of deep gray matter aids in the removal of false positives. Inspection of the bottom of the plots demonstrates the lack of discriminating features associated with the T1 image which is also well-known in the clinical literature.

As described earlier, for Stage 2, we used the output random forest voting maps from Stage 1 as both features themselves and as priors for input to a Bayesian-based segmentation with an additional MRF spatial prior. In

Figure 5, the voting maps are labeled as "*RFStage1VotingMaps*" where the final numeral is associated with the brain parenchymal labeling given previously. Similarly, the additional RF prior segmentation feature probability maps are labeled as "*RFBrainSegmentationPosteriors*". The Stage 2 feature importance plot follows similar trends as that for Stage 1 with the T1 images not contributing much to the identification of white matter hyperintensity voxels. The initial voting maps from Stage 1 are extremely important with the top 3 being the estimated locations of the 1) gray matter, 2) white matter, and 3) white matter hyperintensities. Since these tissue type can be conflated based on intensity alone it is intuitive that such features would be important.

White matter hyperintensity segmentation evaluation

In Figure 7 are the segmentation comparisons derived from manual segmentations of the same data. Despite the large variability characteristic with manual labelings in related fields[36,38,39], such labelings are characteristic of current clinical practices and the methodology proposed herein is readily adapted to refinements in training data.

On the left of Figure 7 are the improvement in Dice values[40], over all white matter hyperintensities when comparing the segmentations between the two stages where the sum is taken over all individually labeled manual, T_r, and automated, S_r, lesions and $\cap$ represents the intersection between the manual/automated lesion pair. Performing the second round of supervised learning improves these Dice values. One can also note from the right side of Figure 6 that the total lesion load volume illustrates a few subjects that are severe outliers in terms of the number of false positives. The second round helps to correct this issue.

Discussion

The current communications describes a supervised statistical learning methodology for identifying WHMs within multimodal MR brain imaging. This effort utilized information acquired from the manual segmentation of WMHs from FLAIR images to help build two-stage ensembles of decision trees for the automated identification of these lesions. Although only a single expert was used to produce the manual labelings, our

intent is to further refine the proposed paradigm by crowdsourcing with feedback from other experts who interact with both the data and methodology. Also, we recognize that only a single site was used for evaluating the proposed framework. However, we are currently processing other site data with the models developed for this work and the results look promising since the developed features are site-agnostic.

As far as we know, this is the first report utilizing a novel random forest approach to identify WMHs in a cohort of TBI patients. TBI WMHs tend to be more difficult to segment than MS lesions as the former tend to be smaller with an overall smaller lesion load. Also, enhancement protocols with the former tend to be less successful than with the latter. As mentioned previously, the work in MS lesion segmentation is extensive with a handful of techniques being publicly available.

Two major meta-analyses of WMHs have been published covering the periods prior to 2010[41] and after[42]. The earlier meta-analysis covered 53 longitudinal studies that included samples of high-risk populations, i.e., patients selected for a specific disease or condition such as hypertension, whereas other studies recruited samples of the general population. Longitudinal studies of samples representative of the general population are more relevant to the focus of the present paper. Debette & Markus[41] found that the presence of WMHs was related to subsequent cognitive decline, a higher risk of developing dementia, stroke, and of mortality. Lesion volume at baseline was also predictive of cognitive decline. Limitations of this meta-analysis include heterogeneity in the method of measuring WMHs; some studies used automated volumetric measurement, whereas others used a visual rating scale. The studies analyzed by Debette & Markus were limited to the occurrence of one of the aforementioned conditions which they analyzed by hazard ratios.

The more recent meta-analysis by Kloppenborg et al.[42] of 23 cross-sectional studies reporting MRI and concurrent neuropsychological results in patients with heterogeneous diagnoses but without previously diagnosed cognitive impairment, found that WMHs were associated with cognitive deficit (effect size of -0.10, 95% CI: -0.13 to -0.08) after controlling for age. These studies also differed in the metric used to measure the WMHs, including volume, % of total intracranial volume, and a visual rating score. The effect size for the association

with cognitive deficit in these cross-sectional studies did not differ significantly across various cognitive domains or the method of measuring lesion volume. Among eight longitudinal studies analyzed by Kloppenborg et al that included a follow-up MRI and also controlled for age, the effect size for the association of progression in WMHs and cognitive impairment was -0.16 (95% CI:-0.27 to -0.09). This association was stronger for attention and executive function than for memory and processing speed. Although baseline WMHs were predictive of cognitive deficit at follow-up in the seven studies which did not repeat MRI, the effect size was smaller [-0.10 (95% CI: 0.13 to -0.05) than in the longitudinal studies that calculated progression in WMHs. In summary, progression of WMHs seen on repeat MRI has a stronger relation to cognitive deficit than concurrent imaging findings. These meta-analyses support the rationale for repeating an MRI in patients younger than 50 years whose initial scan shows WMHs.

Despite the above-described associations between WMHs, cognitive decline, increased risk of developing dementia, and mortality, these lesions receive little attention in current clinical workflows. When reported in a standard neuroradiologist interpretation, they are typically handled as incidental findings and are assigned little clinical significance. This likely reflects the impracticality of performing a detailed assessment of number, volume, and distribution within a qualitative neuroradiologist interpretation as well as the lack of correlative information on how the presence and distribution of these lesions may inform a diagnosis and prognosis in the appropriate clinical setting. To date, automated or semi-automated tools for the detection of WMHs have lacked the specificity and efficiency for the mining of large-scale datasets to generate highly granular data on whether these lesions possess any true diagnostic or prognostic value in the setting of a specific disease process. The present communication describes a supervised statistical learning tool that is appropriate for the application to such large-scale datasets.

The currently described tool is just one example of how “supervised learning” algorithms might be applied to aid in the diagnosis of TBI and other disease processes through the specific identification of features predictive of a given disease state. It is an important demonstration of the potential power of these analytical approaches

in the rapid but comprehensive mining of information from neuroimaging examinations. Supervised learning algorithms are presently employed across a wide variety of settings for the rapid identification of predictive imaging features[43–46]. Automobile manufacturers utilize these types of approaches to equip self-driving vehicles to recognize and respond to unique external surroundings through the identification of visual information sufficiently similar to previously assimilated training data[47,48]. Similarly, in the context of the neuroimaging assessments, deep learning approaches may allow for the rapid identification of information predictive of disease state in an individual patient. These approaches have been applied to the segmentation of macroscopically visible structures[43–46]. Additionally, these approaches might be applied to the interrogation of imaging data in the individual patient with a primary quantitative output metrics to include sequences such as diffusion tensor imaging (DTI) and its variants, functional connectivity, perfusion weighted imaging, and cortical thickness assessments. At present, these advanced neuroimaging sequences are confined to cohort-based research studies due to the lack of available analytical tools to assess the information in the setting of the individual patient[49]. Application of deep learning approaches in the context of data with primary quantitative outputs will require large scale normative and disease specific databases. Building these large scale imaging libraries is resource intensive and requires a multi-center approach with harmonized scanners between sites and correlative non-imaging clinical data. Large scale TBI data is becoming increasingly available through activities such as the Chronic Effects of Neurotrauma Consortium (CENC), Transforming Research and Clinical Knowledge in TBI (TRACK-TBI), Collaborative European Neurotrauma Effectiveness Research in TBI (CENTER-TBI), Department of Defense Alzheimer's Disease Neuroimaging Initiative (DOD-ADNI), and other data being consolidated through FITBIR. In concert with any available high quality normative neuroimaging data, deep learning algorithms may be well positioned to help transform how neuroimaging is interpreted for the clinical management of patients with this disease process.

Acknowledgements

The authors wish to acknowledge all other members of the CENC Neuroimaging Steering Committee and CENC leadership (Drs. David X. Cifu, Ramon Diaz-Arrastia, and Rick Williams) for their support. We also gratefully acknowledge the assistance of Tracy Nolen, Chris Siege and Kevin Wilson. We would also like to thank the study participants and their family members. This project was jointly supported by the Department of Defense (W81XWH-13-2-0095), the U.S. Department of Veterans Affairs (I01 CX001135 and I01 RX 002174), as well as USUHS Grant HU 0001-08-0001.

Declaration of Interest/Disclaimer

The authors report no financial disclosures or conflicts of interest. The views expressed here are those of the authors and do not necessarily reflect the official policy of position of the Department of the Navy, Department of Defense, nor the U.S. Government. This work was prepared as a part of official duties; Title 17 USC §105 provides that Copyright protection under this title is not available for any work of the U.S. Government. Title 17 USC §101 defines a US Government work as a work prepared by a military service member of employee of the US Government as part of that person's official duties.

References

Figure Captions

Figure X: Workflow illustration for the proposed pipeline. Processing of the multi-modal input MRI for a single subject, using the multi-modal symmetric template, results in the generation of the feature images. These feature images are used as input to the Stage 1 RF model producing the initial RF probability map estimates. The Stage 1 voting maps, the original feature images, and the Stage 2 RF model result in the final voting maps which includes the WMH probability estimate. Note that the RF models are constructed once from a set of training data which are processed using the same feature-construction pipeline as the single-subject input MRI.

Figure 1: Representation of Stage 1 feature images for subject 01C1019. The FLAIR, T1-, and T2-weighted images are rigidly pre-aligned[22] to the space of the T1 image. The three modality images are then preprocessed (N4 bias correction[26] and adaptive denoising[25]) followed by application of standard ANTs brain extraction and n-tissue segmentation protocols using the MMRR symmetric template and corresponding priors[28] applied to the T1 image. The feature images are then generated for voxelwise input to the RF model which results in the voting maps illustrated on the right. This gives a probabilistic classification of tissue type. Not shown are the probability and voting images for the brain stem and cerebellum.

Figure 2: Sample FLAIR acquisition image slices showing both manual and random forest segmentations for both stages obtained during the leave-one-out evaluation. Manual segmentations were performed by one of the authors and provided the ground truth WMH labels for training the random forest models.

Figure 3: Canonical views of the mutlivariate, bilaterally symmetric template constructed from the MMRR data set[30] (only shown are the FLAIR, T1, and T2 modalities— the components relevant for this work). Template construction is detailed in[17]. These images are important for specific intensity-based features.

Figure 4: Average *MeanDecreaseAccuracy* plots generated from the creation of all 24 random forest models

for Stage 1 during the leave-one-out evaluation. These plots are useful in providing a quantitative assessment of the predictive importance of each feature. Features are ranked in descending order of importance. The horizontal error bars provide the 95^{th} percentile and illustrate the stability of the feature importance across the leave-one-out models. At this initial stage only 31 feature images are used.

Figure 5: Average *MeanDecreaseAccuracy* plots generated from the creation of all 24 random forest models for Stage 2 during the leave-one-out evaluation. These plots are useful in providing a quantitative assessment of the predictive importance of each feature. Features are ranked in descending order of importance. The horizontal error bars provide the 95^{th} percentile and illustrate the stability of the feature importance across the leave-one-out models. We augment the 31 feature images from the first stage by adding an additional seven voting maps and 7 segmentation posteriors from application of the Bayesian-based segmentation for a total of 45 images for the second stage.

Figure 6: (a) FLAIR image slice illustrating WMHs which have been manually delineated. The region around the WMHs is enlarged (b) in the original FLAIR and the (c) contralateral FLAIR difference image.

Figure 7: Voxelwise comparison with manual delineation of white matter hyperintensities. On the left are the calculated Dice values over all white matter hyperintensities. Note the improvement in the Dice metric from the employment of the Stage 2 component of the processing pipeline. (Right) Similar results can be seen by comparing the total lesion load volume between manual and automated detection strategies. Although some outliers are found after the Stage 2 processing in a couple subjects, the number of outliers caused by false positives is decreased significantly with the second stage processing.

1. Bigler ED, Abildskov TJ, Petrie J, Farrer TJ, Dennis M, Simic N, Taylor HG, Rubin KH, Vannatta K, Gerhardt CA, et al. Heterogeneity of brain lesions in pediatric traumatic brain injury. Neuropsychology. 2013;27(4):438–51.

2. Smitherman E, Hernandez A, Stavinoha PL, Huang R, Kernie SG, Diaz-Arrastia R, Miles DK. Predicting outcome after pediatric traumatic brain injury by early magnetic resonance imaging lesion location and volume. J Neurotrauma. 2016;33(1):35–48.

3. Marquez de la Plata C, Ardelean A, Koovakkattu D, Srinivasan P, Miller A, Phuong V, Harper C, Moore C, Whittemore A, Madden C, et al. Magnetic resonance imaging of diffuse axonal injury: Quantitative assessment of white matter lesion volume. J Neurotrauma. 2007;24(4):591–8.

4. Moen KG, Brezova V, Skandsen T, Håberg AK, Folvik M, Vik A. Traumatic axonal injury: The prognostic value of lesion load in corpus callosum, brain stem, and thalamus in different magnetic resonance imaging sequences. J Neurotrauma. 2014;31(17):1486–96.

5. Ding K, Marquez de la Plata C, Wang JY, Mumphrey M, Moore C, Harper C, Madden CJ, McColl R, Whittemore A, Devous MD, et al. Cerebral atrophy after traumatic white matter injury: Correlation with acute neuroimaging and outcome. J Neurotrauma. 2008;25(12):1433–40.

6. Pierallini A, Pantano P, Fantozzi LM, Bonamini M, Vichi R, Zylberman R, Pisarri F, Colonnese C, Bozzao L. Correlation between mRI findings and long-term outcome in patients with severe brain trauma. Neuroradiology. 2000;42(12):860–7.

7. Weiss N, Galanaud D, Carpentier A, Tezenas de Montcel S, Naccache L, Coriat P, Puybasset L. A combined clinical and mRI approach for outcome assessment of traumatic head injured comatose patients. J Neurol. 2008;255(2):217–23.

8. Levin HS, Williams D, Crofford MJ, High WM Jr, Eisenberg HM, Amparo EG, Guinto FC Jr, Kalisky Z, Handel SF, Goldman AM. Relationship of depth of brain lesions to consciousness and outcome after closed

head injury. J Neurosurg. 1988;69(6):861–6.

9. Breiman L. Random forests. In: Machine learning. 2001. pp. 5–32.

10. Yi Z, Criminisi A, Shotton J, Blake A. Discriminative, semantic segmentation of brain tissue in MR images. Med Image Comput Comput Assist Interv. 2009;12(Pt 2):558–65.

11. Viola P, Jones M, Snow D. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision. 2005;63:153–161.

12. Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage. 2011;57(2):378–90.

13. Pustina D, Coslett HB, Turkeltaub PE, Tustison N, Schwartz MF, Avants B. Automated segmentation of chronic stroke lesions using lINDA: Lesion identification with neighborhood data analysis. Hum Brain Mapp. 2016 Jan.

14. Geremia E, Menze BH, Ayache N. Spatial decision forests for glioma segmentation in multi-channel MR images. In: Proceedings of MICCAI-BRATS 2012. 2012.

15. Bauer S, Fejes T, Slotboom J, Wiest R, Nolte L-P, Reyes M. Segmentation of brain tumor images based on integrated hierarchical classification and regularization. In: Proceedings of MICCAI-BRATS 2012. 2012. pp. 10–13.

16. Zikic D, Glocker B, Konukoglu E, Shotton J, Criminisi A, Ye DH, Demiralp C, Thomas OM, Das T, Jena R, et al. Context-sensitive classification forests for segmentation of brain tumor tissues. In: Proceedings of MICCAI-BRATS 2012. 2012. pp. 1–9.

17. Tustison NJ, Shrinidhi KL, Wintermark M, Durst CR, Kandel BM, Gee JC, Grossman MC, Avants BB. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor seg-

mentation (simplified) with aNTsR. Neuroinformatics. 2015;13(2):209–25.

18. Schapire R. The strength of weak learnability. Machine Learning. 1990;5:197–227.

19. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997;55:119–139.

20. Ho TK. Random decision forests. In: Document analysis and recognition, 1995., proceedings of the third international conference on. Vol. 1. 1995. pp. 278–282 vol.1.

21. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Computation. 1997;9:1545–1588.

22. Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The Insight ToolKit image registration framework. Front Neuroinform. 2014;8:44.

23. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116–28.

24. Neema M, Guss ZD, Stankiewicz JM, Arora A, Healy BC, Bakshi R. Normal findings on brain fluid-attenuated inversion recovery mR images at 3T. AJNR Am J Neuroradiol. 2009;30(5):911–6.

25. Manjón JV, Coupé P, Martí-Bonmatí L, Collins DL, Robles M. Adaptive non-local means denoising of mR images with spatially varying noise levels. J Magn Reson Imaging. 2010;31(1):192–203.

26. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging. 2010;29(6):1310–20.

27. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19(2):143–50.

28. Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, Strien N van, Stone JR, Gee JC, et al. Large-scale evaluation of aNTs and freeSurfer cortical thickness measurements. Neuroimage. 2014;99:166–

79.

29. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics. 2011;9(4):381–400.

30. Landman BA, Huang AJ, Gifford A, Vikram DS, Lim IAL, Farrell JAD, Bogovic JA, Hua J, Chen M, Jarso S, et al. Multi-parametric neuroimaging reproducibility: A 3-T resource study. Neuroimage. 2011;54(4):2854–66.

31. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al. The multimodal brain tumor image segmentation benchmark (bRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024.

32. Maurer CR, Rensheng Q, Raghavan V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2003;25(2):265–270.

33. Anbeek P, Vincken KL, Osch MJP van, Bisschops RHC, Grond J van der. Probabilistic segmentation of white matter lesions in mR imaging. Neuroimage. 2004;21(3):1037–44.

34. Tustison NJ, Avants BB. Explicit B-spline regularization in diffeomorphic image registration. Front Neuroinform. 2013;7:39.

35. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011;54(3):2033–44.

36. García-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med Image Anal. 2013;17(1):1–18.

37. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2/3:18–22.

38. Grimaud J, Lai M, Thorpe J, Adeleine P, Wang L, Barker GJ, Plummer DL, Tofts PS, McDonald WI, Miller

DH. Quantification of mRI lesion load in multiple sclerosis: A comparison of three computer-assisted techniques. Magn Reson Imaging. 1996;14(5):495–505.

39. Styner M, Lee J, Chin B, Chin M, Commowick O, Tran H, Markovic-Plese S, Jewells V, Warfield S, editors. Special Issue on 2008 MICCAI Workshop - MS Lesion Segmentation. MIDAS J; 2008.

40. Tustison NJ, Gee JC. Introducing Dice, Jaccard, and other label overlap measures to ITK. Insight Journal. 2009.

41. Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. BMJ. 2010;341:c3666.

42. Kloppenborg RP, Nederkoorn PJ, Geerlings MI, Berg E van den. Presence and progression of white matter hyperintensities and cognition: A meta-analysis. Neurology. 2014;82(23):2127–38.

43. Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, Johnson HJ, Paulsen JS, Turner JA, Calhoun VD. Deep learning for neuroimaging: A validation study. Front Neurosci. 2014;8:229.

44. Suk H-I, Lee S-W, Shen D, Alzheimer's Disease Neuroimaging Initiative. Deep sparse multi-task learning for feature selection in alzheimer's disease diagnosis. Brain Struct Funct. 2015 May.

45. Li R, Zhang W, Suk H-I, Wang L, Li J, Shen D, Ji S. Deep learning based imaging data completion for improved brain disease diagnosis. Med Image Comput Comput Assist Interv. 2014;17(Pt 3):305–12.

46. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, Feng D, Fulham MJ, ADNI. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. IEEE Trans Biomed Eng. 2015;62(4):1132–40.

47. Hadsell R, Sermanet P, Ben J, Erkan A, Scoffier M, Kavukcuoglu K, Muller U, LeCun Y. Learning long-range vision for autonomous off-road driving. J. Field Robotics. 2009;26(2):120–144.

48. Farabet C, Couprie C, Najman L, LeCun Y. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In: Proceedings of the 29th international conference on machine learning, ICML 2012,

edinburgh, scotland, uK, june 26 - july 1, 2012. icml.cc / Omnipress; 2012.

49. Mayer AR, Bedrick EJ, Ling JM, Toulouse T, Dodd A. Methods for identifying subject-specific abnormalities in neuroimaging data. Hum Brain Mapp. 2014;35(11):5457–70.