Magnetic Resonance Imaging 62 (2019) 18–27

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

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

Original contribution

Automatic identification of atherosclerosis subjects in a heterogeneous MR brain imaging data set

Mariana Bento[a,b,c,*], Roberto Souza[a,b], Marina Salluzzi[a,c], Letícia Rittner[d], Yunyan Zhang[a], Richard Frayne[a,b,c]

[a] Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada
[b] Seaman Family MR Research Centre, Foothills Medical Center, Calgary, AB, Canada
[c] Calgary Image Processing and Analysis Center (CIPAC), Foothills Medical Centre, Calgary, AB, Canada
[d] Medical Image Computing Laboratory (MICLab), School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil

ARTICLE INFO

Keywords:
Brain image processing
Carotid artery atherosclerotic disease
Machine learning
Feature extraction
Multi-center data set

ABSTRACT

Carotid-artery atherosclerosis (CA) contributes significantly to overall morbidity and mortality in ischemic stroke. We propose a machine learning technique to automatically identify subjects with CA from a heterogeneous cohort of magnetic resonance brain images. The cohort includes 190 subjects with CA, white mater hyperintensites of presumed vascular origin or multiple sclerosis, as well as 211 presumed healthy subjects. We determined a set of handcrafted and convolutional discriminant features to perform this task. A support vector machine (SVM) was used to perform this four-class classification task. Our approach had an accuracy rate of 97.5% (higher than chance accuracy of 52.6% for guessing majority class), sensitivity of 96.4% and specificity of 97.9% in identifying subjects with CA, suggesting that the proposed combination of features may be used as an imaging biomarker for characterizing atherosclerotic disease on brain imaging.

1. Introduction

Magnetic resonance (MR) image-based assessment of subjects with carotid-artery atherosclerosis (CA) is commonly dependent on key morphological characteristics, such as the degree of carotid artery stenosis or presence of ulcerated plaques in the artery wall [1]. CA subjects, who are known to be at increased risk of a first or recurrent ischemic stroke [2], commonly present a diffuse pattern of white matter hyperintensities (WMHs) [3]. The association between these brain WMHs and large-vessel atherosclerosis warrants further investigation of these lesions as potentially diagnostic brain imaging features.

MR imaging is a highly suitable technique to non-invasively image the brain. However, manual extraction of white matter information from MR images, particularly in large data sets, is time-consuming and prone to error due to both intra- and inter-operator variability [4]. Thus, more agile and robust computer-assisted diagnosis (CAD) techniques are needed to automatically assess brain features, enhance diagnosis of clinical conditions and improve monitoring of disease progression. Many recent and promising techniques to examine WMHs that are based on machine learning (ML) approaches have demonstrated that automatic methods can produce accurate, quantitative results [5–7].

One of the challenges in developing ML-based algorithms for MR image analysis is related to the potential impact of variability in image acquisition (imaging parameters, pulse sequence implementation, acquisition protocol, software revision, magnetic field strength, scanner vendor, *etc.*), particularly when using a data set acquired at multiple centers. Assessing CAD tool performance using such a multi-center data set, however, is essential for generalizing and validating brain MR measurements and other findings, and is required to more comprehensively evaluate robustness and reliability [8–14].

A key step in a ML-based algorithm is feature extraction. Features extracted from MR brain imaging may represent relevant information about the overall health and disease status of a patient and, potentially, may serve as biomarkers of a disease [3,14,15]. Handcrafted features have traditionally been studied. Such features are defined using *a priori* knowledge about the images and/or disease process [16,17] and are generally specific to an application. As a result while one set of handcrafted features may be appropriate for one disease, they can be inappropriate for other diseases or purposes (segmentation, classification or longitudinal assessment). Convolutional features generated from convolutional neural networks (CNNs), on the other hand, can more generically discriminate regions and patterns within the image [18]. Moreover, combinations of handcrafted and convolutional features may

* Corresponding author at: Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada.
E-mail address: mariana.pinheirobent@ucalgary.ca (M. Bento).

https://doi.org/10.1016/j.mri.2019.06.007
Received 13 February 2019; Received in revised form 4 June 2019; Accepted 6 June 2019

lead to improved results with enhanced understanding and interpretation [19].

Our goal is to extract, evaluate and identify features that allow for the automatic identification of subjects with CA from a diverse cohort of MR brain images. Included in this cohort were subjects with 1) observed WMHs due to a variety of other etiologies, as well as 2) presumed healthy subjects. A multi-class classifier was developed by combining handcrafted features (principally texture-based features [20]) with convolutional features (obtained using a transfer learning approach [21]). We sought to find sets of features that may be used as biomarkers to identify carotid-artery atherosclerosis disease.

The robustness of the method was examined using single- and multi-center data sets, and by considering subjects with a variety of WMH etiologies and both low and high WMH burden. We anticipated some variation by center due to utilization of different scanners with similar but not the same image acquisition parameters. We also anticipated that the WMH burden would have an impact on the classifier. It is expected that subjects with lower WMH burden would be harder to correctly classify because fewer abnormal brain voxels would be present in the image.

2. Multi-center data set

A key requirement when developing ML methods is to have access to robust, generalizable and correctly annotated data sets in order to perform experiments and to validate approaches. Our experiments involved secondary use of four data sets, containing images from both patients and healthy controls. These images were collected at eight centers and contained subjects from four classes (labelled as CA = carotid-artery atherosclerosis, MS = multiple sclerosis, SVD = small vessel disease, and NC = normal controls) that are described below and summarized in Table 1). In order to provide a pragmatic assessment of our proposed method, we trained and tested on a large but unbalanced multi-center, multi-disease data set. Secondary analyses (described below) were used to understand the impacts of potential confounding factors due to this choice of data set.

Across all classes, two-dimensional T2-weighted fluid attenuated inversion recovery (FLAIR) images were acquired over the whole brain in a predominantly axial orientation (*i.e.*, parallel to the anterior commissure-posterior commissure line). Clinically, the FLAIR imaging sequence is commonly used to evaluate brain tissue abnormalities, such as WMHs [7]. Timing parameters including repetition time (TR), echo time (TE), and inversion time (TI), as well as slice thickness and acquisition matrix varied slightly by class and sometimes within each class. Data were collected from 401 subjects using 3-T MR scanners from three vendors.

All subjects in the patients cohort and many from the NC group presented with WMHs. The fraction of subjects with visible WMHs was 252/401 (62.8%). The WMHs were identified and segmented using a semi-automatic software tool (CEREBRA-WML [22]) that uses a region-growing method. The WMH burden (volume) was sub-labelled as being low or high using an empiric threshold of 2,000 voxels [23] (corresponding to a threshold varying from 2 ml to 5 ml, depending on the acquired image resolution). This voxel threshold was chosen over a volume threshold for implementation reasons. In the pre-processing step (described below), all images were resized to have the same matrix size ($48 \times 48 \times 48$ voxels), so the dichotomization between low and high WMH volume ensured that WMH voxels have a similar presentation in the pre-processed images.

Table 1
Overview of the data set demographics including center, number of subjects, average age (mean ± standard deviation), age range (minimum, maximum), biological sex (percent female) and fraction of MR scans with WMH: CA = carotid-artery atherosclerosis (subdivided by recruiting center), MS = multiple sclerosis, SVD = small vessel disease (subdivided by recruiting center), and NC = normal control. Due to secondary use, detailed age and sex information was only known for some data. n/a = not available; WMH = white matter hyperintensity.

Data	Center	No subjects	Age (years)	Age range (years)	Sex (% female)	WMH fraction
CA	1	19	n/a	n/a	n/a	100.0%
	2	22	75 ± 7	(63,94)	59%	100.0%
	3	17	n/a	n/a	n/a	100.0%
	4	41	n/a	n/a	n/a	100.0%
	5	12	n/a	n/a	n/a	100.0%
	Subtotal	111	n/a	n/a	n/a	100.0%
MS	2	19	48 ± 13	(28,75)	100%	100.0%
SVD	6	20	n/a	n/a	n/a	100.0%
	7	20	n/a	n/a	n/a	100.0%
	8	20	n/a	n/a	n/a	100.0%
	Subtotal	60	n/a	n/a	n/a	100.0%
NC	2	211	55 ± 22	(20,87)	61%	29.4%
All		401				62.8%

2.1. Class 1 - carotid artery atherosclerostic disease: CA

Subjects with CA disease were drawn from five study centers (centers 1–5) participating in the Canadian Atherosclerosis Imaging Network (CAIN) Project 1 [24]. Images were acquired from 111 subjects (Fig. 1a–e). Three study centers (1, 2, and 4) used the same 3-T MR scanner platform (Discovery 750; General Electric Healthcare, Waukesha, WI) with TR/TE/TI = 9700 ms/145 ms/2200 ms. Images from center 3 (3-T Achieva; Philips Medical Systems, Best, The Netherlands) used a similar acquisition sequence with parameters TR/TE/TI = 9000 ms/125 ms/2800 ms. Center 5 used a scanner from a third vendor (3-T Tim Trio; Siemens Health Engineering, Erlangen, Germany) also with a comparable acquisition sequence and parameters (TR/TE/TI = 9000 ms/119 ms/2500 ms). All subjects in this class had WMHs.

2.2. Class 2 - multiple sclerosis: MS

Subjects with MS were drawn from a single-center database (center 2 with images acquired using TR/TE/TI = 6000 ms/130 ms/1855 ms) [25], and served as one positive (*i.e.*, diseased) control group. This class includes 19 subjects diagnosed with a range of MS severity. The data included heterogeneous pathology severity as measured by the Expanded Disability Status Scale (EDSS) score. This class included nine severely (EDSS score of ≤ 3) and ten mildly disabled (EDSS score > 6) subjects (Fig. 1f–h). All subjects in this class had WMHs.

2.3. Class 3 - small vessel disease: SVD

Sixty subjects with SVD that had WMH of presumed vascular origin [26] were included. These subjects were obtained from a publicly available data set (WMH Segmentation Challenge held at the 2017 International Conference on Medical Image Computing and Computer Assisted Intervention [27] and served as a second positive control (disease) group. Data were acquired from 20 subjects at three different centers (centers 6 to 8) using different 3-T scanners and similar acquisition parameters (center 6: Signa HDxt, General Electric Healthcare with TR/TE/TI = 8000 ms/126 ms/2340 ms; 7: Achieva with 11,000 ms/125 ms/2800 ms; and 8: Tim Trio with 9000 ms/82 ms/2500 ms; Fig. 1i–k). The age and gender of this data set was not reported. All subjects in this class had WMHs.

2.4. Class 4 - normal controls: NC

Presumed healthy subjects were drawn from the Calgary Normative Study database (center 2) [28]. This is an ongoing study that recruits healthy participants into a longitudinal study. In this study, only baseline images were used and participants with significant incidental findings on MR imaging or with scores < 26 (out of 30) on the Montreal

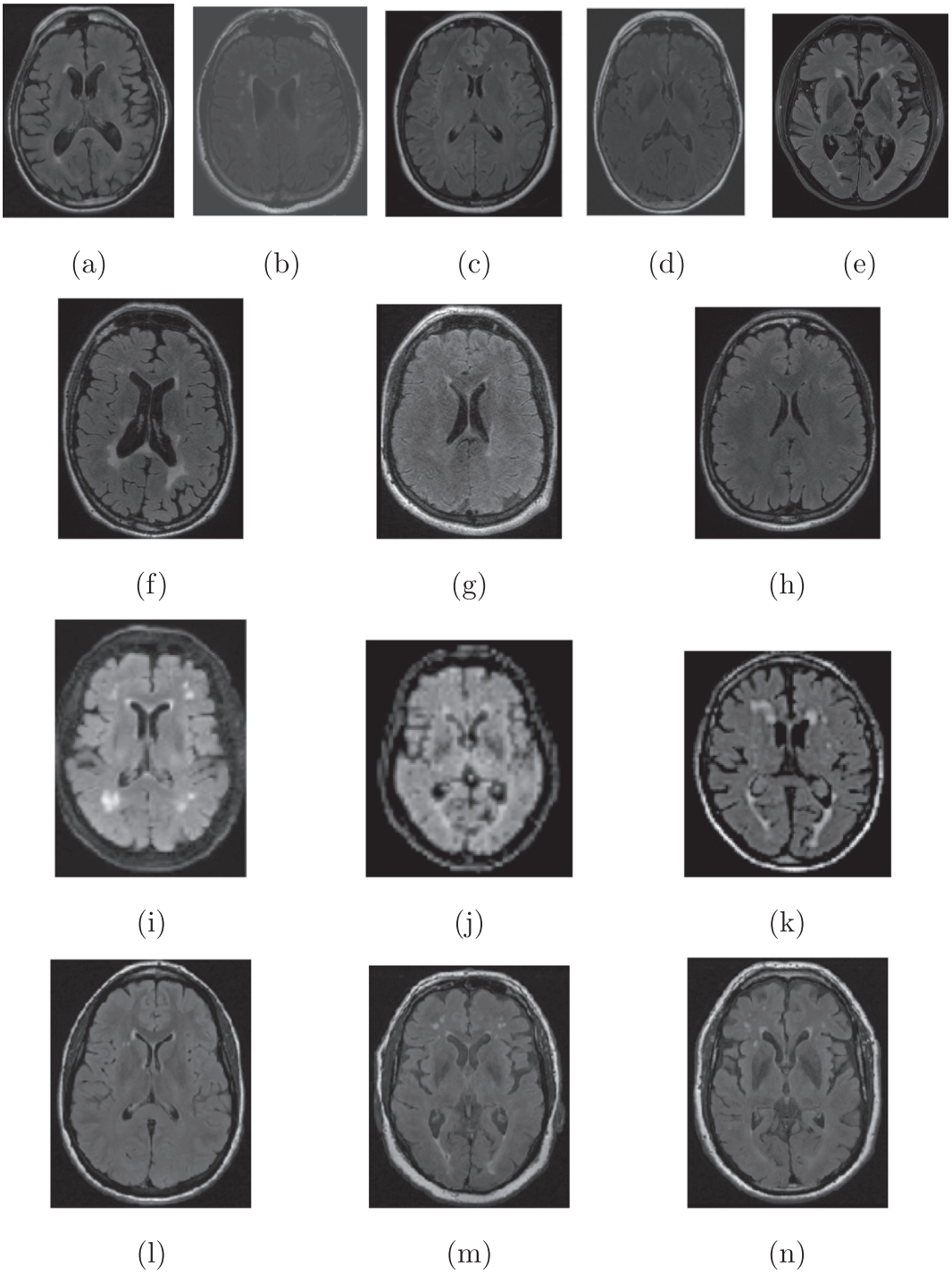


Fig. 1. Example of FLAIR images in the multi-center data sets for each class: (a–e) CA from centers 1 to 5, respectively; (f–h) MS from center 2; (i–k) SVD from centers 6 to 8, respectively; and (l–n) NC from center 2.

Cognitive Assessment (MoCA) test were excluded (Fig. 1l–n). From this database, images from 211 subjects served as a negative (*i.e.*, presumed disease-free) control group (normal controls). In this class, 29% (62 of 211) of the subjects presented with WMH; a finding that is not unexpected in a normative brain aging study [29]. Inclusion of these subjects made the classification task challenging and realistic.

3. Methods

In order to identify CA subjects from positive (MS, SVD) and negative (NC) control subjects, we implemented a four-class classification method consisting of four principal steps: 1) pre-processing, 2) feature extraction, 3) classification and 4) evaluation of results using hand-crafted and convolutional features (see flowchart in Fig. 2).

3.1. Pre-processing

Step 1 included image-intensity correction, brain extraction, image cropping and intensity normalization [30] (Fig. 3). Images acquired at different centers (and on different scanners) with variable acquisition parameters can present inter-center/scanner contrast and intensity variation (*cf.*, Fig. 1). Thus, pre-processing aimed to standardize theMR

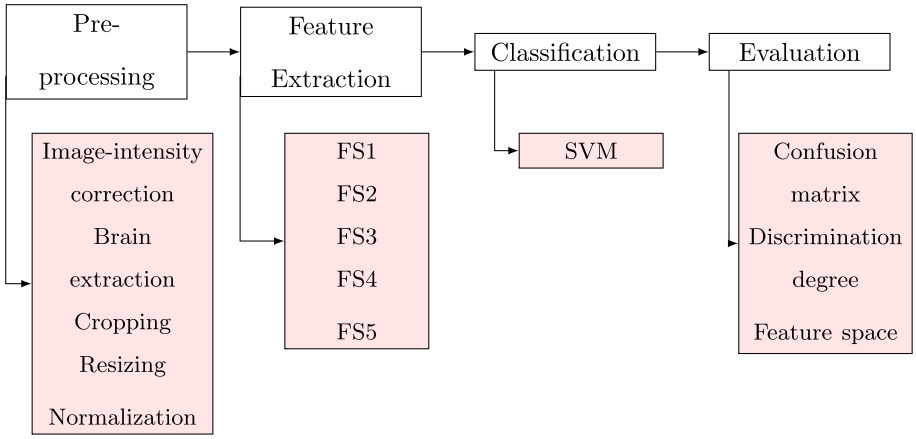


Fig. 2. Main steps of the classification algorithm: 1) pre-processing, 2) feature extraction, 3) classification, and 4) evaluation. FS = two-dimensional feature space, SVM = support vector machine.

images to ensure that distinctions between classes are related to the presence/absence of pathology, the type of pathology and the pattern of the WMH burden, and are not simply related to inherent image variability.

First, the N4 correction method [31] was applied to compensate for non-uniformity across the images by performing bias field correction [32]. Then, the brain was extracted using an automated pipeline (brain extraction tool (BET) from FSL [33]). The third step cropped the image to only include the segmented brain. The last step was image resizing to 48 × 48 × 48 voxels and image intensity normalization to the range [0,255].

3.2. Feature extraction

After pre-processing, we performed feature extraction. This step is important because one of the goals of this work was to find and evaluate sets of features that best identified CA subjects. Our approach was to implement, explore and finally rank different methods to extract features from the images using both traditional [34] and convolutional approaches [35]. We ran multiple experiments using different features sets (FS), in order to evaluate their degree of discrimination in the proposed classification task. The predefined FS were:

- **FS1 - image intensity-based features** from the gray level image histogram. Statistical moments extracted from the histogram provide a quantitative analysis of the image intensity distribution [36]. Eight separate features were extracted from the histogram: [34] entropy, intensity mean, standard deviation, skewness, kurtosis, and values at 10, 50 (median) and 90 percentiles.
- **FS2 - image gradient-based features**. The gradient examines directional changes in the image gray levels, and can extract relevant information (such as edges) within an image. Moments extracted from the image gradient are more robust to acquisition conditions, such as contrast variation, and properties of the acquisition equipment [36]. Ten features were extracted from the gray level and morphological gradients (five from each): [34] intensity mean, standard deviation, skewness, kurtosis and percentage of non-zero values [7].
- **FS3 - features extracted from the local binary pattern (LBP)**. LBP is a texture spectrum model that may be used to identify patterns in an image [37]. The LBP histogram comprises the frequency of occurrence of different patterns within an image [34]. Ten features were extracted from the LBP by using a 10-bin LBP histogram.
- **FS4 - frequency domain features**. The Haar wavelet is a multiresolution technique that transforms images into a domain where both spatial and frequency information is present. Features separately extracted from each sub-image present desired scale-dependent properties [38]. When considering two decomposition levels, eight sub-images are generated [34]. The mean value within each sub-image were computed and used as features (total of eight features).
- **FS5 - convolutional network features**. These were extracted from a very deep convolutional network (VGG16) [39] with pre-trained imagenet weights through transfer learning. These features are expected to be representative and discriminative to the proposed classification task [21,35,40]. For each MR volume, the convolutional features were computed in the central two-dimensional (2D) axial, sagittal and coronal slices. For each of these three views, 25,088 convolutional features were computed, generating a vector of 75,264 features per image.

A total of 75,300 features were extracted from each image: 8

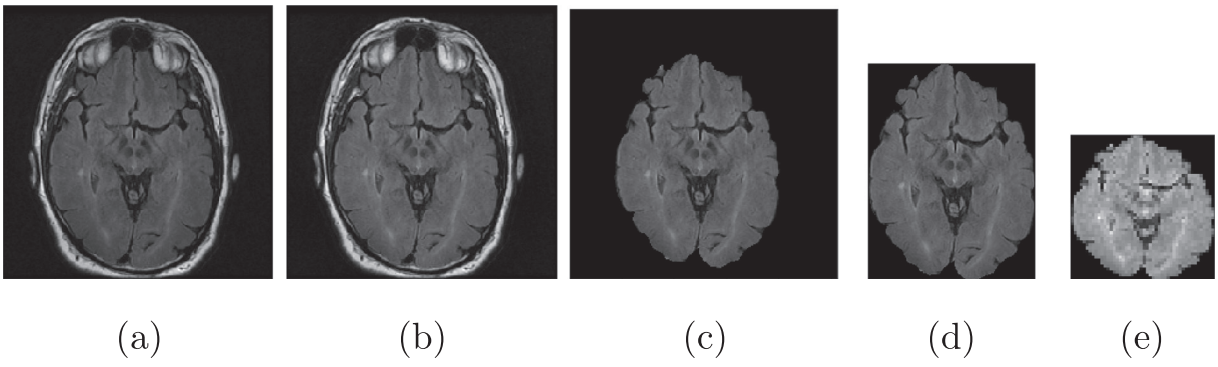


Fig. 3. Pre-processing steps: (a) original image, (b) N4 corrected, (c) brain extracted, (d) brain cropped, and (e) resized and normalized image.

features from the image histogram (FS1), 10 features from the image gradient (FS2), 10 features from the LBP histogram (FS3), 8 features from the Haar wavelet sub-images (FS4) and 75,264 features from the convolutional networks (FS5). These features were used to develop the classifier model to distinguish CA subjects from subjects in the positive (MS, SVD) and negative (NC) control groups. The individual sets (FS1, FS2, FS3, FS4, and FS5) as well as the union of the handcrafted features (FS1 + FS2 + FS3 + FS4) and the union of all features (FS1 + FS2 + FS3 + FS4 + FS5) were examined for the same classification task.

3.3. Classification

Classification was performed to distinguish images from subjects into the four classes (CA, MS, SVD, and NC) by using the previously described features. A support vector machine (SVM) was the chosen classifier model, because it has previously provided good results in similar studies [41] and because SVM is not sensitive to redundant/non-discriminant features [42]. This last finding is important because one of the main goals of this study was to find a set of features that could be used as a biomarker of carotid atherosclerotic disease. The optimal SVM parameters were determined by using a grid search technique [43] only considering the training data. A ridge regularization (L2 regularization) was used because the number of features was large in comparison with the number of subjects [44].

3.4. Evaluation

The performance of the proposed classifier for distinguishing the four classes was evaluated by using a stratified ten-fold cross validation procedure [45]. Similar to traditional cross validation, at each iteration, the data was randomly partitioned into ten folds, one fold for testing, the other folds for training. Because we had an unbalanced data set, we used stratified cross validation, which preserved the class proportions in each fold. The average confusion matrix and other quantitative measures, such as accuracy rate, sensitivity and specificity, over ten iterations were calculated.

We also evaluated the effect of each FS by using quantitative and qualitative approaches including computing their degree of discrimination and visualizing the 2D feature space. The degree of discrimination was computed by using the mean decrease impurity (or Gini importance) method [46] using a decision tree structure [47]. The feature space was visualized by using a 2D reduction of the data set onto a plane defined by the two eigenvectors corresponding to the two largest principal components. These components were computed using the principal component analysis algorithm [48].

We performed additional experiments to analyze the impact of the extracted FS, to check the degree of discrimination for individual groups of features, and to assess if the FS were complementary or redundant. The convolutional features (FS5) were also subjectively analyzed by using a 2D visualization. In this visualization, the learned convolutional weights in a shallow and deeper layer are shown as image slices in a mosaic. This analysis allowed further subjective interpretation of the extracted convolutional features.

We explored the effects of center (over centers 1–5) and WMH burden (low *vs* high) on the proposed 4-class classification task. We performed an additional analysis using CA subjects images (class of interest) acquired at each of the single centers and compared these results to the results achieved across all centers. These additional experiments were performed without changing the number of subjects or centers in the other three classes (MS, SVD and NC). To further assess the effect of center, we conducted a second classification of the image data based on center. The effect of WMH burden was assessed by examining the impact on classification accuracy when using low or high WMH burden samples.

3.5. Statistical methods

One-way analysis of variance (ANOVA) [49] was used to determine significance in the accuracy rate by FS and center. Performance difference among the handcrafted (FS1, FS2, FS3, FS4, FS1 + FS2 + FS3 + FS4), convolutional (FS5) and combined (FS1 + FS2 + FS3 + FS4 + FS5) FS were examined using paired two-sample *t*-tests with Bonferroni correction [50]. A paired two-sample *t*-test was used to examine effects of WMH burden. We did not explore the interaction between center and WMH burden, because of small subgroup sizes. A chi-squared test [51] was used to compare the results between the proposed classification task (based on the disease) and a classification based on center to evaluate the effect of center variability in the proposed method. A p-value < 0.05 was used to determine significance in all tests. Mean ± standard deviation values were reported.

3.6. Implementation environment

Experiments were conducted using Python language (version 2.7.1) and Jupyter Notebook environment [52] on a Mac Pro with a 2.7GHz 12-Core Intel Xeon processor with 64GB DDR3 memory. We used the following common software packages to perform some of the processing: scikits-image (skimage) [53] for traditional feature extraction, Keras [54] for extracting the convolutional features, scikits-learning (sklearn) [42] for the classification models and evaluation metrics, and matplotlib [55] for visualization purposes.

4. Results

The proposed classification method using a combination of handcrafted and convolutional features achieved an average accuracy rate of 97.5 % ± 1.9% (mean ± standard deviation) when performing a four class classification task to discriminate CA subjects from positive control subjects (MS and SVD) and from negative control subjects (NC) (Fig. 4). The chance accuracy for this task, defined as the accuracy to guess all samples as the majority class, was 52.6% (only NC subjects properly classified = 211/401 subjects). Ten subjects were incorrectly labelled (Fig. 4(a)): Two CA subjects were labelled as SVD, and two as NC. Three SVD and one NC subjects were labelled as *CA*. One SVD subject was labelled as NC, and one NC subjects as SVD. Because we have unbalanced data, we also computed sensitivity and specificity for our class of interest. Our sensitivity was 96.4% (107/111 CA subjects), and specificity of 97.9% (284/290 non CA subjects).

Qualitative visualization of the principal components of the 2D feature space demonstrated discrimination of the four class clusters (Fig. 4(b)), confirming the confusion matrix results and the high measured accuracy rate, sensitivity and specificity. Overlap was found between the CA, SVD and NC groups, with the MS group being most distinct. It is important to highlight that this two-dimensional visualization is a simplification of a higher-dimensional space that allows for better separation between the classes.

Analysis of the degree of feature discrimination (Fig. 4(c)) showed that the majority of the most discriminant features were convolutional based. One handcrafted feature ranked within the top thirty discriminant features. This handcrafted feature was based on the LBP approach (in FS3), which comprises the frequency of occurrence of different patterns within the analyzed image.

4.1. Effect of feature set

A high accuracy rate (97.3 % ± 2.1%, Table 2) was achieved when using only convolutional features (FS5), similar to the accuracy achieved when using all features (FS1 + FS2 + FS3 + FS4 + FS5, 97.5 ± 1.9%). A one-way ANOVA across the five FS (FS1, FS2, FS3, FS4 or FS5), the union of the handcrafted features (FS1 + FS2 + FS3 + FS4) and the union of all features

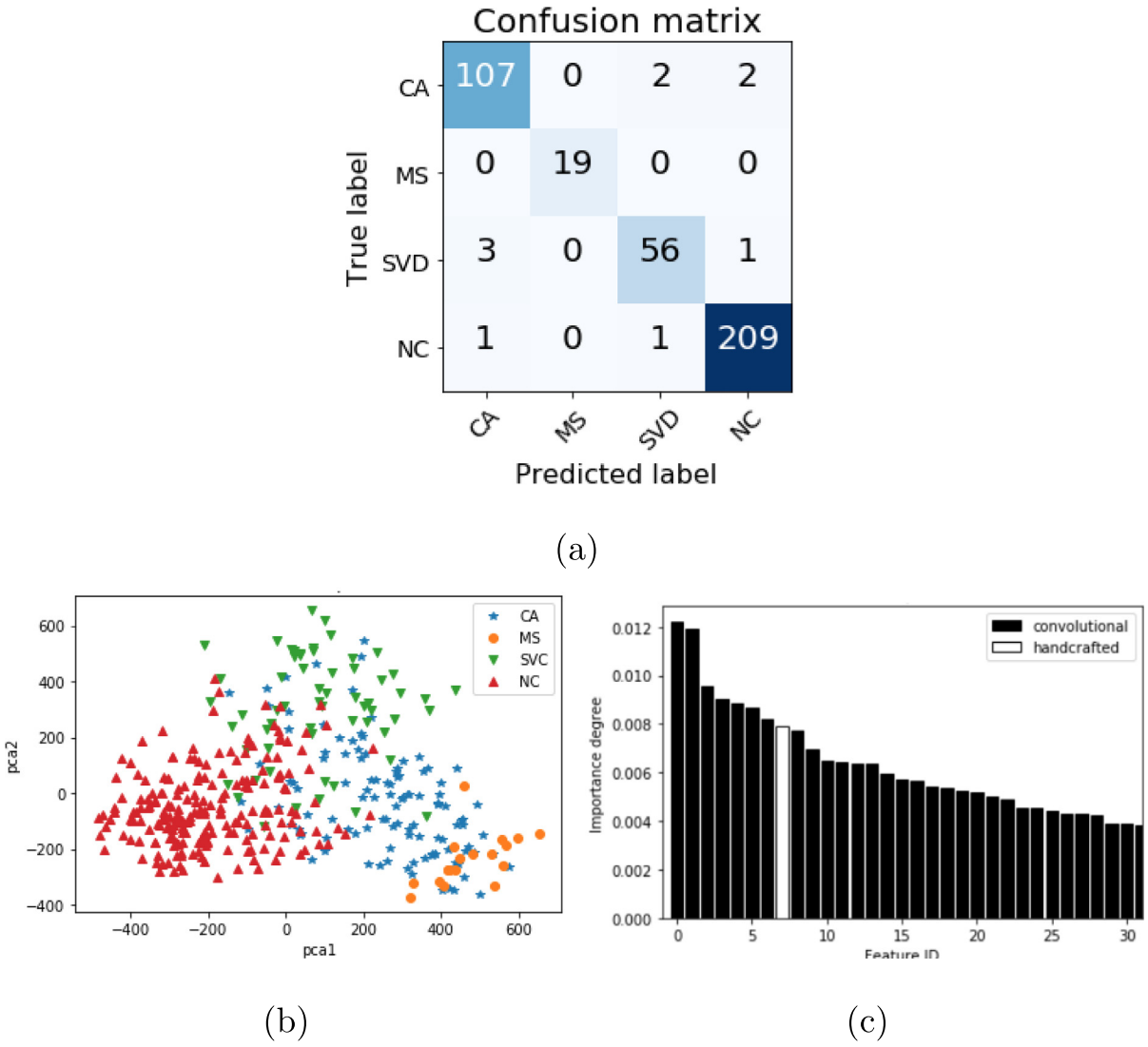


Fig. 4. Results of distinguishing four different classes carotid-artery atherosclerosis (CA), multiple sclerosis (MS), small vessel disease (SVD) and normal control (NC). Shown are (a) confusion matrix, (b) qualitative analysis in a two-dimensional feature space (obtained using principal component analysis, see text), and (c) feature quantitative analysis obtained by computing the feature importance/discrimination degree for the proposed task using a decision tree algorithm (for simplicity, only the 30 most discriminating features are shown, see text).

Table 2
Accuracy rates (mean ± standard deviation) achieved when distinguishing CA from MS, SVD and NC by using handcrafted features (FS1, FS2, FS3, FS4, FS1 + FS2 + FS3 + FS4), or convolutional features (FS5) or a combination of all features sets (FS1 + FS2 + FS3 + FS4 + FS5). Similar accuracies were achieved when combining all FS, or when using only convolutional ones (FS5). A one-way ANOVA on individual features, union of handcrafted features (FS1 + FS2 + FS3 + FS4) and combination of all features was significant ($p < 0.001$). Displayed p-value is from paired t-tests relative to FS1 + FS2 + FS3 + FS4 + FS5. Bonferroni corrected level of significance due to multiple comparison was $p < 0.008$.

Features	Accuracy (%)	p-Value
FS1	90.0 ± 2.8	< 0.001
FS2	89.3 ± 3.8	< 0.001
FS3	75.1 ± 4.6	< 0.001
FS4	69.9 ± 4.3	< 0.001
FS1 + FS2 + FS3 + FS4	94.3 ± 3.1	0.017
FS5	97.3 ± 2.1	0.803
FS1 + FS2 + FS3 + FS4 + FS5	97.5 ± 1.9	

(FS1 + FS2 + FS3 + FS4 + FS5) confirmed the presence of significant differences ($p < 0.001$). Pair-wise significant differences were found between FS1 + FS2 + FS3 + FS4 + FS5 and each of FS1, FS2, FS3, FS4. After Bonferroni correction, the accuracy of the union of the handcrafted features (FS1 + FS2 + FS3 + FS4) and the convolutional features (FS5) were not significantly different than FS1 + FS2 + FS3 + FS4 + FS5 (see Table 2 for p-values).

To better understand the convolutional features (FS5), we visualized these most discriminating features using a 2D visualization mosaic. The computed convolutional features (convolutional weights) comprised a mixture of low and high spatial frequency information at varying orientations. The shallow layers contain simple features and the deep layers presented more complex features (Fig. 5).

Additional analysis of the FS was performed by evaluating the confusion matrices using only handcrafted (FS1 + FS2 + FS3 + FS4) or convolutional features (FS5) (Fig. 6). Because misclassifications occurred in different classes, this suggested that these FS were non-redundant, further supporting the usage of hybrid methods that combine handcrafted and convolutional features.

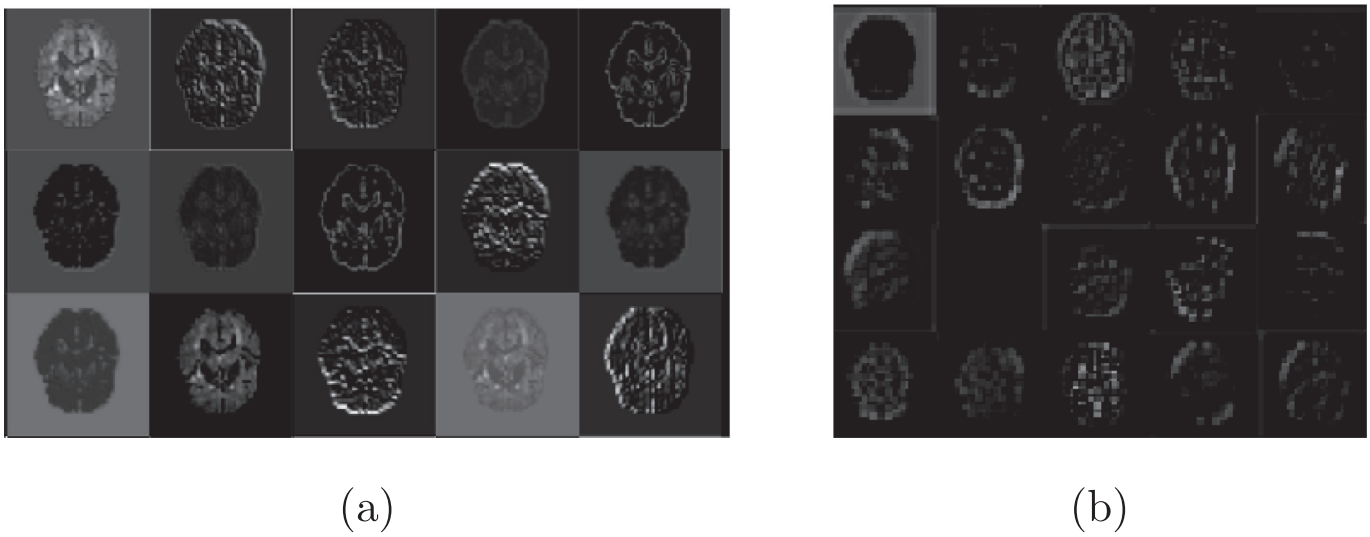


Fig. 5. Visualization of convolutional features (*i.e.*, learned convolutional weights) as a two-dimensional slice mosaic after two VGG16 convolutional layers: (a) initial layer ("block1_pool layer" [39]) and (b) a deeper layer ("block3_pool layer"). The network automatically detects low- and high-level features. The deeper layer presents more complex features.

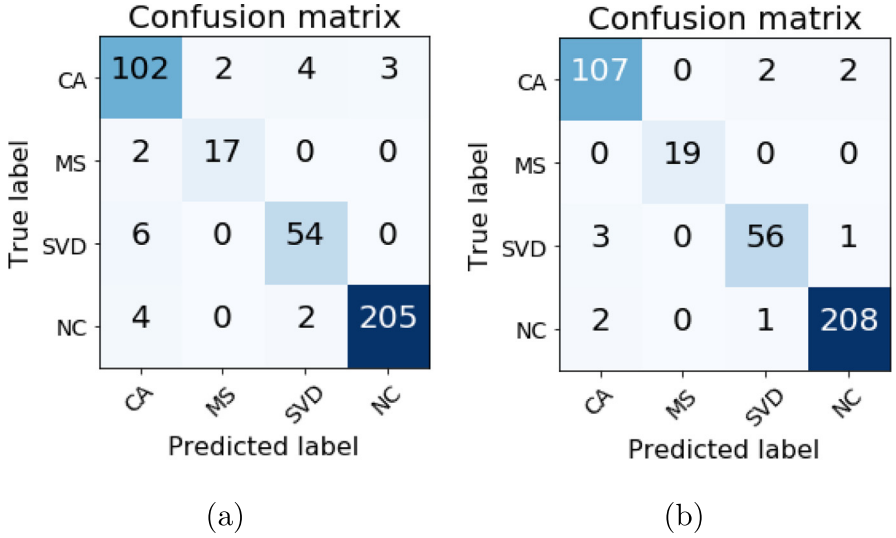
4.2. Effect of center and WMH burden

Small, not significant ($p = 0.127$, one-way ANOVA) variations (Table 3) in accuracy were observed when using the CA data from each single center to perform our proposed four-class classification (*i.e.*, to distinguish CA subjects from MS, SVD or NC). We expected some variation by center due to utilization of different scanners and small changes in the image acquisition parameters.

In order to further assess the effect of center, we performed a second classification of the image data based on center (Fig. 7). This center-wise classification achieved an accuracy of 88.6 % ± 3.6%. Erroneous classificationswere concentrated across centers 1–5, especially center 2, that contain images from patients with different diagnosis, confirming imaging variability due to disease (variability that it is detected by our proposed four-class disease-based classification method). No classification errors were observed for centers 6–8. While the center classification had an acceptable accuracy rate, the accuracy achieved by the proposed four-class disease-based classification was significantly higher (97.5%, $p = 0.00157$, chi-squared test to compare accuracy's achieved

Table 3
Support vector machine (SVM) accuracy rates (mean ± standard deviation) in performing a four class classification task to identifying carotid-artery atherosclerosis subjects with varying data sets: single *vs* multi center, low *vs* high WMH volume. Separate one-way ANOVA tests showed that classification scores were not significantly different by center ($p = 0.127$) or by WMH volume ($p = 0.834$).

Center	WMH volume	No. CA subjects	Accuracy (%)
Multi-center	All	111	97.5 ± 1.9
Center 1 only	All	19	96.5 ± 2.7
Center 2 only	All	22	98.4 ± 1.6
Center 3 only	All	17	96.8 ± 5.0
Center 4 only	All	41	98.8 ± 2.0
Center 5 only	All	12	96.0 ± 2.9
Multi-center	Low	72	97.8 ± 2.1
Multi-center	High	39	97.0 ± 3.3



(a) (b)

Fig. 6. Confusion matrices by varying FS used in the classifier: (a) only using handcrafted features (FS1 + FS2 + FS3 + FS4) - accuracy of **94.3** % ± **3.1**%, and (b) only using convolutional features (FS5) - accuracy of **97.3** % ± **2.1**%.

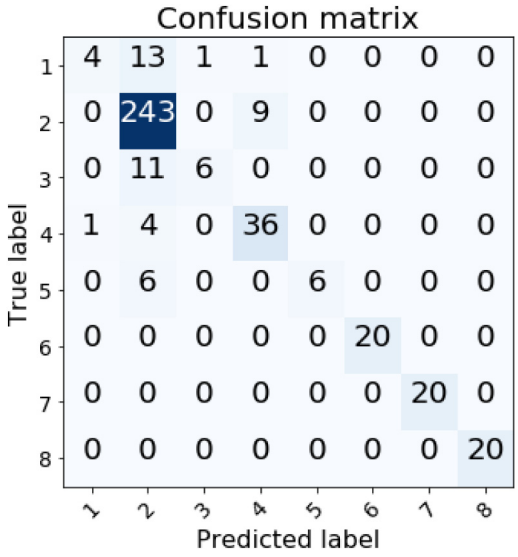


Fig. 7. Results for distinguishing samples based on the center they were acquired (eight class classification task) by using same features and classifier (SVM with cross-validation) used to distinguish images based on the pathology (CA, MS, SVD or NC). Overall accuracy rate to classify center was **88.6** % ± **3.6**%.

to perform classification by center or disease-based classification).

We expected that subjects with lower WMH burden (Fig. 8) would be more difficult to correctly classify. Our results (Table 3), however, showed no significant difference between low and high WMH burden groups (p = 0.834, paired t-test); 72 of 111 CA subjects were found to have low WMH burden.

5. Discussion

Our classifier was able to identify CA subjects from a large, unbalanced, group of MS, SVD and NC subjects with a high accuracy rate (97.5%). In addition to calculating the confusion matrix and accuracy rate, we also evaluated the FS1 + FS2 + FS3 + FS4 + FS5 space by using principal component analysis technique (Fig. 4(b)). Some overlap between the samples in this 2D representation was expected because this visualization is a simplification of the original high-dimensional feature space. The selected classification model (SVM) is known to be appropriate for performing classification of complex data (with overlap samples between classes) [41].

Comparison against other classification studies is not possible because to the best of our knowledge, there are no other works that propose similar classification tasks. However, studies that included multi-center data, as proposed here, are frequently reported as being reliable and more generalizable, presenting superior results when evaluated in data different from the training data [23,56]. In the absence of comparative studies, we instead chose to perform additional experiments to more comprehensively evaluate the impact of confounding variables (such as center) on the performance of our method. We varied the FS used by the classifier, as well as the data set (*e.g.*, single centers *vs* multi-center), and explored the impact of WHM burden. We also examined the false negative and positive findings.

The degree of feature discrimination (Fig. 4(c)) showed that the majority (29/30, 96.7%) of the thirty most discriminant features were convolutional based. This finding might have been expected because of the large difference in the number of handcrafted (36) *versus* convolutional (75,364) features. Despite this numerical imbalance, one handcrafted feature ranked in the top thirty discriminant features, suggesting that independent features are being identified by handcrafted and convolutional methods.

The overall confusion matrix (Fig. 4(a)) showed that most of the false negative and false positive misclassifications were found between classes CA and NC and between CA and SVD. One possible explanation for these errors is that the abnormalities, such as the WMHs, observed in CA, SVD and 29% of the NC data (see Table 1) are most likely of similar origin, ischemic [26]. In contrast, MS subjects have WMHs principally associated with a demyelination [7]. This difference in etiology likely impacts the observed 2D feature space visualization as more overlap occurred among the CA, SVD and NC groups than with the MS group (Fig. 4(b)).

Our data was unbalanced, with one of our classes (MS) having only 19 subjects. MS is a rarer condition than carotid artery or small vessel disease, so the imbalance reflects disease prevalence. MS subjects are known to have heterogeneous image appearance, related to the presence and severity of the pathology. This variation may impact the MR imaging characteristics of an individual. However, while it is a small sample, this study did include subjects with disabilities ranging from very mild to severe. Nonetheless, additional study using a larger MS group would be beneficial. Conversely NC, our largest class, contains subjects, having a range of WMH burdens. Even with this variability in the negative-control (*i.e.*, disease free) class, high disease classification accuracy rates were achieved. The use of unbalanced data from a variety of centers, suggests robustness and potential generalizability of our approach.

The combination of handcrafted and convolutional features

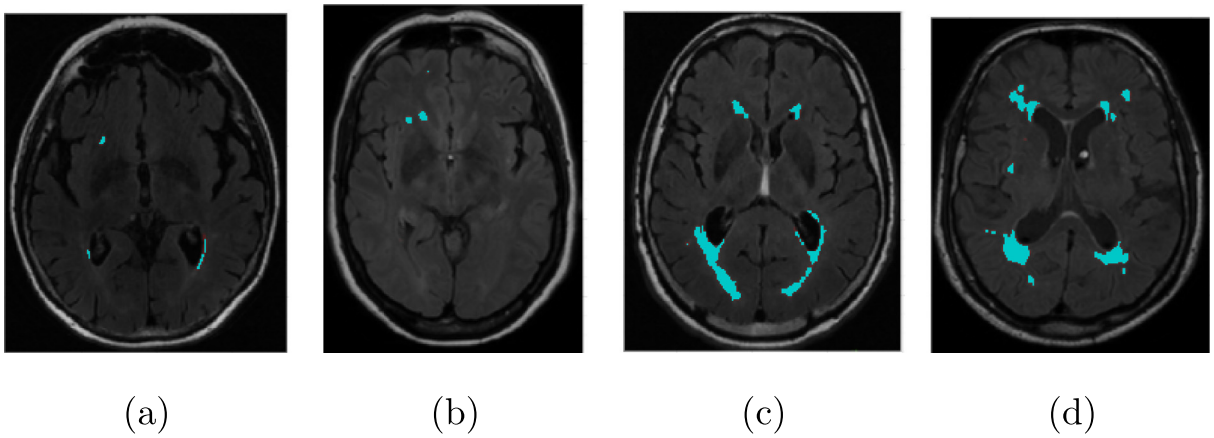


Fig. 8. WMH burden (light blue areas) in four sample images from the CA data set. WHM volume was dichotomized using a 2000 voxel threshold (corresponding to volumes of between 2 ml and 5 ml, depending on the acquired image resolution) into: (a, b) low and (c, d) high WMH volume groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(FS1 + FS2 + FS3 + FS4 + FS5) could identify subjects with CA by distinguishing them from positive and negative controls (proposed four-class classification task) and possibly serves as an imaging biomarker of disease presence.

5.1. Effect of feature set

We used in this study a combination of handcrafted (FS1, FS2, FS3 and FS4) and convolutional features (FS5). The number of handcrafted features were limited (36); the inclusion of more features, such as the ones extracted from the gray level co-occurrence and run length matrices [57] might improve the results given by the handcrafted features sets. In contrast, the convolutional methods extracted 75,264 features.

Interpretation of convolutional features is currently an active area of research and remains one of the biggest challenges in using deep convolutional networks in medical imaging problems, because such models are commonly considered to be a black box, showing no audit trail to explain their decisions [58]. A visualization of these features was presented (Fig. 5). This visualization allows a subjective analysis of these features in an image space, highlighting some anatomical regions and characteristics that were considered relevant to performing the four-class classification.

The combination of handcrafted features (FS1 + FS2 + FS3 + FS4) presented an average accuracy of 94.3% (see Table 2). An accuracy of 97.3% was achieved when using the convolutional features (FS5). In analysis systems with limited amount of computational and storage resources, the combination of handcrafted features approach could be used, with only a slight reduction in accuracy.

5.2. Effect of center and WMH burden

Center and WMH burden were found to have no significant impact on the experiments. Small variations ($p = 0.127$) in accuracy rates were observed in the single-center analysis of the CA data when performing our proposed four-class classification (distinguishing subjects from classes CA, MS, SVD or NC). Two centers (center 2 and 4) had higher accuracy, while three centers (centers 1, 3 and 5) presented lower accuracy rates compared to the multi-center analysis. The observed variations, however, were small, not significant, and may simply reflect that the centers with the highest accuracy had the largest cumulative enrollment (centers 2 and 4, 63/111 subjects, 56.8%).

Further experiments were performed to study the influence of the imaging variability in the proposed model by classifying images based on the center they were acquired. The comparison between the results achieved when classifying images based on the pathology or based on the acquisition center showed that: 1) the accuracy of these tasks are statistical different; and 2) even though our proposed model presented a high accuracy rate (almost 10% higher), there is likely some influence of the imaging variability when distinguishing CA subjects from MS, SVD and NC. The pre-processing techniques that were applied to standardize the images and handle the variability of the data related with the acquisition center reduced but could not eliminate all of this variability. Further exploration of image standardization to remove center characteristics might alleviate the influence of the center on the classification task.

Similar accuracy rates were achieved for the low (97.8%) and high WMH volumes (97.0%). Both accuracy rates were not statistically different to that found with the entire data set (97.5%), suggesting some robustness of our method to the WMH burden in classifying subject pathology. A possible reason is that we use the whole brain in our classification and not smaller regions of interest containing only WMH and adjacent tissue.

6. Conclusions

Distinguishing normal from pathological images within a large and complex, unbalanced, multi-center data set is challenging, even when performed by imaging specialists. Here, we presented an automatic method that distinguishes carotid-artery atherosclerosis (CA) subjects from subjects with other neurodegenerative diseases (MS and SVD, positive controls) and from healthy subjects (NC, negative controls). Our focus was on exploring both traditional (handcrafted) and CNN-based (convolutional) feature extraction methods.

A 97.5% accuracy rate was achieved for the four-label classification problem (*i.e.*, CA, MS, SVD *vs* NC). No significant statistical changes in accuracy were observed when performing the four class classification task using CA subjects subgroups based on center and WMH volume. Significant differences, however, were observed by FS.

Further experiments assessing the effects of the center in the classification (classification by center) showed that our results may be influenced by inter-center variability, despite applying image pre-processing techniques to minimize these effects. However, we demonstrated that our proposed model to classify images based on the subject pathology is statistically different than a model that just distinguishes images based on acquisition center. Similar results also were achieved when using single center CA data or when using multi-center CA data. Thus, our model may be suitable for images acquired at multiple centers using different scanners with similar but not the same acquisition parameters.

We also evaluated the degree of discrimination for individual features sets. The observed accuracy rate using only convolutional features (FS5) was similar to the accuracy achieved when combining all handcrafted features. However, we found non-redundancy between the handcrafted and convolutional features. The model performed best when using a combination of handcrafted and convolutional features (hybrid features). This combination may be used as a biomarker for the identification of carotid-artery atherosclerosis in FLAIR images of the brain.

In future work, additional evaluation with multi-center MS and NC data should be used. This additional study will allow experiments to further evaluate and improve the proposed model robustness to image/center variability. Additional data could also be segregated and used as a distinct testing data set for evaluation of future models. We also intend to work on a semi-adversarial classification model that considers final accuracy as a combination of accuracies that distinguish the images based on the diseases and center. Our goal in such an approach would be to have a high accuracy to distinguish disease, while penalizing high accuracy for distinguishing images based on center (in order to minimize the effects of center). Finally, we also intend to explore our FS, and develop more robust features to handle the imaging variability, not related with specific brain abnormalities, such as WMH burden.

Declaration of Competing Interest

The authors have no conflict of interest to declare.

Acknowledgments

We would like to thank various groups for the secondary use of their data sets: 1) The Canadian Atherosclerosis Imaging Network (CAIN Principal Investigator: Jean-Claude Tardif, MD; and CAIN Project 1 lead: Alan Moody, MD). CAIN was funded by the Canadian Institutes for Health Research (CIHR), the Canada Foundation for Innovation (CFI) and the Governments of Alberta, Ontario and Québec. 2) The MS Society of Canada for funding the MS image acquisition (Principal Investigator: Lenora Brown, PhD) and the Calgary MS Clinic for supporting the study. 3) The organizers of the WMH Segmentation Challenge (2017 International Conference on Medical Image Computing and Computer Assisted Intervention [27]; 4) The NC data was drawn from the Calgary Normative Study and this analysis was supported by the CIHR (MOP-333931; Principal Investigator: Richard Frayne, PhD).

Mariana Bento, PhD was supported by the Hotchkiss Brain Institute (HBI) Burns Fellowship. Roberto Souza, PhD was supported by the T Chen Fong Postdoctoral Fellowship. Leticia Rittner, PhD thanks the National Council for Scientific and Technological Development (CNPq, Brazil; 308311/2016-7). Richard Frayne, PhD is supported by the Hopewell Professorship in Brain Imaging. The Calgary-Campinas collaboration was supported in part by an award from the Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil), Special Visiting Professor Program (PVE-88881.062158/2014-01).

References

[1] Saba L, Yuan C, Hatsukami TS, Balu N, Qiao Y, DeMarco JK, et al. Carotid artery wall imaging: perspective and guidelines from the ASNR vessel wall imaging study group and expert consensus recommendations of the American Society of Neuroradiology. Am J Neuroradiol 2018:1–23.
[2] Flaherty ML, Kissela B, Khoury JC, Alwell K, Moomaw CJ, Woo D, et al. Carotid artery stenosis as a cause of stroke. Neuroepidemiology 2012;40(1):36–41.
[3] Ammirati E, Moroni F, Magnoni M, Rocca M, Messina R, Anzalone N, et al. Relation between characteristics of carotid atherosclerotic plaques and brain white matter hyperintensities in asymptomatic patients. Sci Rep 2017;7(10559):1–11.
[4] Leite M, Rittner L, Lapa A, Appenzeler S, Lotufo R. Classification of brain white matter lesion on MR imaging. Proceedings of the 15th Annual Alberta Biomedical Engineering Conference. 2014.
[5] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42(Supplement C):60–88.
[6] Leite M, Gobbi D, Salluzi M, Frayne R, Lotufo R, Rittner L. 3D texture-based classification applied on brain white matter lesions on MR images. Proceedings Volume 9785: Medical Imaging 2016: Computer-aided Diagnosis SPIE. 2016.
[7] Leite M, Rittner L, Appenzeller S, Ruocco H, Lotufo R. Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. Journal of Medical Imaging 2015;2(1):014002. [1–014002: 10].
[8] Tofts P, Collins D. Multicentre imaging measurements for oncology and in the brain. Br J Radiol 2011;84:213–26.
[9] Chalavi S, Simmons A, Dijkstra H, Barker G, Reinders A. Quantitative and qualitative assessment of structural magnetic resonance imaging data in a two-center study. BMC Med Imaging 2012;12(27):1–15.
[10] Davids M, Zollner F, Ruttorf M, Nees F, Flor H, Schumann G, et al. Fully-automated quality assurance in multi-center studies using MRI phantom measurements. Magn Reson Imaging 2014;32(6):771–80.
[11] Kim E, Magnotta V, Liu D, Johnson H. Stable atlas-based mapped prior (STAMP) machine-learning segmentation for multicenter large-scale MRI data. Magn Reson Imaging 2014;32(7):832–44.
[12] Nelson A, Swallen A, Chung S, Piper J. Evaluation of a quantitative metric, volumetric statistical amyloid burden (VSAB), for Florbetapir PET, to classify amyloid positive and negative subjects using a cutoff derived from an independent dataset. The Journal of Nuclear Medicine 2016;57(2):551.
[13] Heinen R, Bouvy W, Mendrik A, Viergever M, Biessels G, Bresser J. Robustness of automated methods for brain volume measurements across different MRI field strengths. Plos One 2016;11(10):e0165719.
[14] Nieuwenhuis M, Schnack H, Haren N, Lappin J, Morgan C, Reinders A, et al. Multi-center MRI prediction models: predicting sex and illness course in first episode psychosis patients. Neuroimage 2017;145(Pt B):246–53.
[15] Moeskops P, de Bresser J, Kuijf HJ, Mendrik AM, Biessels GJ, Pluim JP, et al. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in mri. NeuroImage: Clinical 2018;17:251–62.
[16] Loizou C, Pantziaris M, Pattichis C, Seimenis I. Brain MR image normalization in texture analysis of multiple sclerosis. Journal of Biomedical Graphics and Computing 2013;3(1):20.
[17] Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 2004;22(1):81–91.
[18] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.
[19] Kashif M, Raza S, Sirinukunwattana K, Arif M, Rajpoot N. Handcrafted features with convolutional neural networks for detection of tumor cells in histology images. Proceedings of IEEE 13th International Symposium on Biomedical Imaging. 2016.
[20] Castellano G, Bonilha L, Cendes F. Texture analysis of medical images. Clin Radiol 2004;59(12):1061–9.
[21] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35(5):1285–98.
[22] Lu Q, Gobbi DRF, Salluzzi M. Cerebra-WML: a stand-alone application for quantification of white matter lesion. Proceedings of Imaging Network Ontario Symposium. 2014.
[23] Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, et al. BIANCA (Brain Intensity Abnormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. NeuroImage 2016;141:191–205.
[24] Tardif J, Spence J, Heinonen T, Moody A, Pressacco J, Frayne R, et al. Atherosclerosis imaging and the Canadian Atherosclerosis Imaging Network. Can J Cardiol 2013;29(3):297–303.
[25] Komishke B, Metz LM, Brown LN, Zhang Y. Preferably large fiber damage in the corpus callosum of progressive MS compared with relapsing MS and controls. Proceedings of the 22nd International Society for Magnetic Resonance in Medicine. 2014.
[26] Wardlaw J, Smith E, Biessels G, Cordonnier C, Fazekas F, Frayne R, et al. Neuroimaging standards for research into small vessel disease and its contribution to aging and neurodegeneration. Lancet Neurol 2013;12:822–38.
[27] Kuijf HJ, Biesbroek JM, de Bresser J, Heinen R, Andermatt S, Bento M, et al. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the wmh segmentation challenge. IEEE Trans Med Imaging 2019. https://doi.org/10.1109/TMI.2019.2905770.
[28] Tsang A, Lebel CA, Bray SL, Goodyear BG, Hafeez M, Sotero RC, et al. White matter structural connectivity is not correlated to cortical resting-state functional connectivity over the healthy adult lifespan. Front Aging Neurosci 2017;9(144):1–13.
[29] Kim K, MacFall J, Payne M. Classification of white matter lesions on magnetic resonance imaging in the elderly. Biol Psychiatry 2009;64(1):273–80.
[30] Leite M, Rittner L, Gobbi D, Saluzzi M, Frayne R, Lotufo R. Influence of MR image intensity normalization on texture-based classification of brain white matter lesions. Proceedings of the Second Brain Congress. 2015.
[31] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–20.
[32] Belaroussi B, Milles J, Carme S, Zhu YM, Benoit-Cattin H. Intensity nonuniformity correction in MRI: existing methods and their validation. Med Image Anal 2006;10(2):234–46.
[33] Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. Neuroimage 2012;62:782–90.
[34] Bento M, Souza R, Frayne R, Rittner L. WMH segmentation challenge: a texture-based classification approach. Lecture Notes in Computer Science: International MICCAI Brain Lesion Workshop. 2017.
[35] Bento M, Souza R, Salluzzi M, Frayne R. Normal brain aging: prediction of age, sex and white matter hyperintensities using a MR image-based machine learning technique. Lecture Notes in Computer Science: International Conference Image Analysis and Recognition. 2018.
[36] Woods R, Gonzalez RC. Digital image processing. Edgard Blucher; 2000.
[37] He DC, Wang L. Texture unit, texture spectrum, and texture analysis. IEEE Transactions on Geoscience and Remote Sensing 1990;28(1):509–12.
[38] Sandeep C, Patnaik L, Jagannathan N. Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. Biomedical Signal Processing and Control 2006;1(1):86–92.
[39] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of 3rd International Conference on Learning Representations. 2015.
[40] Bento M, Souto L, Salluzzi M, Zhang Y, Frayne R. Feature extraction using convolutional networks for identifying carotid artery atherosclerosis patients in a heterogeneous brain MR dataset. Proceedings of Joint Annual Meeting of International Society for Magnetic Resonance in Medicine. 2018.
[41] Yichuan T. Deep learning using linear support vector machines. Proceedings of International Conference on Machine Learning. 2013.
[42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011;12(1):2825–30.
[43] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal Machine Learning Research 2012;13(1):281–305.
[44] Bishop CM. Pattern recognition and machine learning. Springer; 2007.
[45] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence. Vol. 2. USA: San Francisco; 1995. p. 1137–43.
[46] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1 ed. Wadsworth International Group; 1984.
[47] Han J, Kamber M. Data mining: concepts and techniques. 2 ed. Elsevier; 2006.
[48] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 1901;2(11):301.
[49] Winkler R, Hays W. Statistics: probability, inference, and decision. 2st ed. Holt, Rinehart and Winston, Inc.; 1975.
[50] Dunnett CW. A multiple comparisons procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50(272):1096–121.
[51] Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 1940;11(1):86–92.
[52] Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and power in academic publishing: players, agents and agendas. 2016. p. 87–90.
[53] der Walt S, Schonberger JL, Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: image processing in python. PeerJ the Journal of Life and Environmental Sciences 2014;2:e453.
[54] Chollet F, et al. Keras. https://github.com/fchollet/keras; 2015.
[55] Hunter JD. Matplotlib: A 2d graphics environment. Computing In Science & Engineering 2007;9(3):90–5. https://doi.org/10.1109/MCSE.2007.55.
[56] Helmer K, Chou M, Preciado R, Gimi B, N. R, Song A, et al. Multi-site study of diffusion metric variability: effects of site, vendor, field strength, and echo time on regions-of-interest and histogram-bin analyses. Proceedings of SPIE–the International Society for Optical Engineering. 2016.
[57] Schwartz WR, Siqueira FR, Pedrini H. Evaluation of feature descriptors for texture classification. Journal of Eletronic Imaging 2012;21(2):1–17.
[58] Yamashita R, Nishio M, Do R, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights into Imaging 2018;9(4):611–29.