Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – A systematic review

Ramya Balakrishnan a, c, 1, Maria del C. Valdés Hernández a, b, 1, *, Andrew J. Farrall a, b

a Edinburgh Imaging Academy, University of Edinburgh, Edinburgh, UK
b Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
c International Institute of Health Sciences, Sri Lanka

ARTICLE INFO

Keywords:
White matter lesions
White matter hyperintensities
Supervised segmentation
Unsupervised segmentation
Deep learning
FLAIR hyperintensities

ABSTRACT

Background: White matter hyperintensities (WMH), of presumed vascular origin, are visible and quantifiable neuroradiological markers of brain parenchymal change. These changes may range from damage secondary to inflammation and other neurological conditions, through to healthy ageing. Fully automatic WMH quantification methods are promising, but still, traditional semi-automatic methods seem to be preferred in clinical research. We systematically reviewed the literature for fully automatic methods developed in the last five years, to assess what are considered state-of-the-art techniques, as well as trends in the analysis of WMH of presumed vascular origin.

Method: We registered the systematic review protocol with the International Prospective Register of Systematic Reviews (PROSPERO), registration number - CRD42019132200. We conducted the search for fully automatic methods developed from 2015 to July 2020 on Medline, Science direct, IEE Explore, and Web of Science. We assessed risk of bias and applicability of the studies using QUADAS 2.

Results: The search yielded 2327 papers after removing 104 duplicates. After screening titles, abstracts and full text, 37 were selected for detailed analysis. Of these, 16 proposed a supervised segmentation method, 10 proposed an unsupervised segmentation method, and 11 proposed a deep learning segmentation method. Average DSC values ranged from 0.538 to 0.91, being the highest value obtained from an unsupervised segmentation method. Only four studies validated their method in longitudinal samples, and eight performed an additional validation using clinical parameters. Only 8/37 studies made available their methods in public repositories.

Conclusions: We found no evidence that favours deep learning methods over the more established k-NN, linear regression and unsupervised methods in this task. Data and code availability, bias in study design and ground truth generation influence the wider validation and applicability of these methods in clinical research.

1. Introduction

In 1987, Hachinski, Potter, and Merskey (Hachinski et al., 1987) first used the term leukoaraois to describe abnormal areas of decreased density in subcortical white matter on brain computed tomography (CT) scans. Leukoaraois has also been referred to as white matter lesions (WMLs) (Inzitari, 2003). With increasing use of magnetic resonance imaging (MRI) as a diagnostic tool, leukoaraois is increasingly referred to as white matter hyperintensities (WMH) (Wardlaw et al., 2013).

Being one of the most studied neuroimaging features given their appearance in a large number of pathologies and in normal ageing, the term WMH is indistinctively used to refer to abnormal clusters of T2-weighted-based hyperintense signal in tissue, usually larger than 3 mm diameter, which are not artificially induced by the imaging system (Wardlaw et al., 2013). WMHs are associated with reduced cognitive function, dementia, gait, balance, mobility, and mood disorders (Fazekas and Wardlaw, 2013; Zheng et al., 2011). WMHs are also frequently observed in the asymptomatic aged and associated with common...
geriatric conditions such as cerebrovascular disease, cardiovascular disease, multiple sclerosis, other autoimmune diseases and psychiatric disorders such as depressive disorder, bipolar disorder and schizophrenia (Kim et al., 2008; Rachmadi et al., 2018). WMH prevalence in the general population ranges from 11 to 21% in 64 year olds and increases with age to 94% in 82 year olds (Debette and Markus, 2010). One study reported that amongst an elderly population aged 60–90 years, 90% have WMH (Hasan et al., 2019).

Detailed WMH evaluation for number, volume, location, and distribution on MRI may provide crucial information on aetiology, prognosis, and progression of diseases; accurate quantification may help measure treatment effectiveness (Manjón et al., 2018; Qin et al., 2018). WMH severity is considered an indirect marker of normal appearing white matter integrity and a surrogate marker of small vessel disease (SVD) (Malais et al., 2019; Maniago et al., 2015). Advancing MRI technology means several methods have been developed to quantify WMH volumes through image segmentation: “a process which typically partitions the spatial domain of an image into mutually exclusive subsets called regions, each one of which is uniform and homogeneous with respect to some property such as tone or texture and whose property value differs in some significant way from the property value of each neighbouring regions” (Haralick and Shapiro, 1991). However, WMH are not homogeneous, have ill-defined boundaries and their tone and texture may not significantly differ from neighbouring tissues. Biologically, they represent the “tip of the iceberg” of demyelinating, inflammatory processes which affect the whole brain: they accompany and sometimes coalesce with many neuroradiological features. Essential for digital image segmentation is recognition of edges which separate WMH from “background”. WMH identification subjectivity and boundary recognition, challenge WMH segmentation, leading to low agreement in studies of manual delineation of WMH ground truth segmentations (Akudjedu et al., 2018; Despotovic et al., 2015; Keller and Roberts, 2009).

Unlike normal tissues, for which validated fully automatic protocols exist and have become standard, WMH segmentation is, albeit mature, an active field of research for which a myriad of methodologies are still being developed. Clinical research groups usually select a WMH segmentation method based on their own capabilities, existing methods’ specifications, availability and sustainability of the source code, and image acquisition protocols. Then, groups adapt these methods in-house and validate them for a specific study protocol. Normal tissue intensities follow a normal distribution, but abnormalities do not. In the specific case of WMH, signal intensity and spatial distributions vary, displaying unique signatures for each disease and cohort. Table 1 summarises some WMH signatures in normal ageing, SVD, Alzheimer’s disease (AD), multiple sclerosis (MS) and vanishing white matter disease (an autosomal recessive disorder) (Labaigue et al., 2009). In addition to specific disease / neurological condition characteristics, WMH appearances vary widely in individuals from different disease groups (Fig. 1).

WMHs arising as a result of infections (e.g. viral, bacterial), can overlay those which already exist due to other processes (e.g., normal ageing) or comorbidities (e.g., SVD): this poses a challenge for their differential identification and segmentation. For example, in COVID-19 patients, in addition to large vessel strokes, WMHs have been reported bilaterally in the thalami, cerebellum and temporal lobes, and also in the corpus callosum, along with abnormal T2 signal in the olfactory bulb and microbleeds in the thalami (Imaging in COVID-19 complications -)

Table 1

| Neuroradiological signatures of T2 WMHs in normal ageing, SVD, AD, MA and vanishing WMD. |
|------------------|-----------------|-----------------|-----------------|-----------------|
| Normal ageing    | SVD             | Alzheimer’s disease | Multiple sclerosis | Vanishing white matter disease |
| Extent           | From thin peri-ventricular lining to confluent deep WM regions | From peri-ventricular with few deep WM foci to confluent regions | Thin peri-ventricular lining with foci in deep WM | From peri-ventricular with few deep WM foci to large confluent regions (may enclose “pseudo-cavities” of low T1 signal, also referred as “cavitory lesions” (Ayrignac et al., 2016) |
| Characteristic brain regions | Non-specific. Disseminated throughout periventricular & deep WM & corpus striatum. Gradual progression extending from periventricular WM, from frontal & parietal regions. Rare in temporal lobes, brain stem & cerebellum. | Non-specific. Disseminated throughout periventricular & deep WM, corpus striatum & thalami. Rare in cerebellum. | Non-specific. Disseminated throughout the whole brain, including midbrain, brainstem, juxtacortical regions, corpus callosum & cerebellar peduncles. | Disseminated throughout the whole brain. External capsule & corpus callosum diffusely involved. |
| Symmetry between brain hemispheres | Symmetric distribution | Symmetric distribution | Symmetric distribution in cerebrum, but not in cerebellum | Symmetric distribution in cerebellum |
| Histogram distribution in FLAIR MRI | Tail (from normal WM) fits. | Tail (from normal WM) fits. | Very skewed independent of (i.e. separated from) that of normal WM. | Bimodal independent of (i.e. separated from) that of normal WM (considering cavitation). |
| Signatures | “Cotton-wool” appearance. | “Cotton-wool” appearance combined with subtel hyperintense large areas, or overt & confluent with irregular patterns. | “Cotton-wool” appearance. Confluent regions are uncommon. | Lesions juxtacortically & in corpus callosum, perpendicular to ventricular linings (known as “Dawson fingers”). Predominantly periventricular cavitory lesions. Punctate asymmetrical posterior fossa lesions. |
| Presence of other features | Perivascular spaces (from none to mild) & possibly lacunar lesions // chronic mild ischaemic strokes. | Perivascular spaces, lacunes, chronic ischaemic strokes & few microbleeds. | Marked brain atrophy & hippocampal sclerosis. | Diffuse corpus callosum involvement. Symmetric involvement of cerebellum & middle cerebellar peduncles. |
| Borders | Not well defined. | Not well defined. | Better defined than in SVD or healthy ageing. | Not well defined. |

(Griffanti et al., 2016)
ESR Connect, 2020). In these brain regions, typical WMHs are uncommon. Symmetric frontal WMH and cortical hyperintensities have been reported in other COVID-19 patients with more severe respiratory disease status (MRI Shows Brain Abnormalities in Some COVID-19 Patients, 2020) along with punctate cortical blooming artefacts. But influence of treatment, comorbidities and disease severity make it difficult even for neuroradiologists to identify specific disease-related patterns that could differentially aid in diagnosis and patient stratification.

To help select WMH segmentation methods and discuss their applicability, other systematic literature reviews have been published, but on focused topics specific to diseases, e.g. MS lesion segmentation (García-Lorenzo et al., 2013; Lladó et al., 2011, 2012; Miller et al., 1998; Mortazavi et al., 2012). Methods which work for MS may only perform moderately if applied to individuals with SVD or to the normal elderly (Table 1). Caligiuri et al. conducted a systematic review on fully automated methods for segmenting WMH in normal ageing and in patients with vascular pathology and risk factors, covering from 1980 to 2014 (Caligiuri et al., 2015). Two other non-overlapping reviews (Wardlaw et al., 2015; Blair et al., 2017) discussed different approaches published up to 2016, both for segmenting WMH, and also for assessing other neuroimaging markers of SVD. Another study which systematically reviewed machine-learning methods which differentiate healthy aging from different dementia types (Pellegrini et al., 2018) included studies (from 2006 to September 2016) aimed at detecting and segmenting WMH in ageing and dementia. The last five years (i.e., since 2015) have seen a boost in sample sizes, computational power and the introduction / application of deep learning in clinical research, in parallel with an increase in high-quality imaging acquisitions, facilitated by 3 T MRI scanners.

We systematically reviewed the literature from 2015 to 2020 in order to assess and overview those fully automatic computational methods developed to segment WMH of presumed vascular origin.

2. Methods

2.1. Literature search

This systematic review protocol is registered on the International Prospective Register of Systematic Reviews (PROSPERO), registration number - CRD42019132200 (2020) to avoid unintended duplication.
and to aid in transparent reporting. The search was conducted from January 2015 to July 2020 on Medline, Science direct, IEE Explore and Web of Science. For each database, we developed a search strategy to retrieve as many WMH segmentation method articles as possible. We identified keywords by expanding the subject components from the review question: white matter lesion, white matter hyperintensities, leukoaraiosis, aging, WMH, segmentation, supervised segmentation, unsupervised segmentation, machine learning, deep learning, parcellation, artificial neural network, pattern recognition, clustering, classification, magnetic resonance imaging, MRI. We applied language restriction and age limits (45 plus years) for Medline. We summarize search strategy details for each database in Appendix 1. We imported all articles retrieved into the reference manager Mendeley, and removed all duplicates. We then screened abstracts and titles to exclude studies outside the scope of the review. Then we evaluated the full text of the remaining articles, applying inclusion and exclusion criteria (explained below). We also reviewed references of these articles for possible papers missed in the primary search.

Additionally, the following journals were hand-searched to identify articles which presented a method for segmenting WMH in the period covered by this review.

1. NeuroImage – keywords WMH and segmentation from January 2015 to July 2020: identified 6 articles which matched the search results (100 % recall).
2. NeuroImage Clinical – from January 2015 to July 2020: identified 7 articles which matched the search results (100 % recall).
3. Neuroinformatics – keywords WMH and segmentation from January 2015 to July 2020: identified 2 articles which matched the search results (100 % recall).

2.2. Inclusion and exclusion criteria

2.2.1. Inclusion criteria

- Presentation and / or validation of a fully automated method for segmenting WMH of presumed vascular origin from human brain MR images from January 2015 to July 2020
- Studies published in English

2.2.2. Exclusion criteria

- Animal studies
- Segmentation methods solely for MS lesions or validated using only MS patients
- New fully automated segmentation methods not described
- Segmentation method accuracy evaluated on CT images
- Segmentation of brain regions, brain tumours or other pathologies which are not WMH of presumed vascular origin
- Semi-automated segmentation methods
- No information on segmentation method similarity metrics used or no evaluation against ground truth segmentations
- Focus on pre-processing or classification of MR images
- Methods designed for WMH evolution assessment (i.e. longitudinal analysis) without cross-sectional validation of their results
- Insufficient information to replicate or apply the segmentation method
- Studies published only as abstracts
- Conference proceedings
- No segmentation method description

2.3. Assessment of methodological quality

We evaluated methodological quality for each study using QUADAS 2: a tool to assess the risk of bias and the applicability of the methods / procedures (https://www.bristol.ac.uk/media-library/sites/quadas/mi grated/documents/quadas2.pdf). QUADAS 2 contains four domains: 1) patient selection; 2) index test; 3) reference test; and 4) flow and timing. In our case, index text refers to WMH segmentation method / algorithm. Different from the original QUADAS 2 questionnaire, the evaluation of the index text consisted in assessing whether or not the reference standard was used in any way by the segmentation method. We completed the online form for each of the included studies. If a study were judged low in all four domains in relation to bias or applicability from answering the specific questions from each domain, then it was considered as “low risk of bias”. If a study were judged high or unclear for one or more domains, then it was considered as “risk of bias” or as having concerns regarding applicability.

2.4. Data extraction

From the included papers, we extracted the following data:

- Title, year of publication, journal name, study design
- Number of subjects or images, age, gender
- Patient selection criteria, sample size
- Type of MRI sequences used (details about the scanner used)
- Information on imaging features used for investigation
- Details about pre-processing steps (registration, brain extraction, intensity inhomogeneity correction, noise reduction, intensity normalization)
- Method to remove false positives
- Reference standard(s)
- Segmentation method details
- Non-imaging features used for clinical correlation with WMH volume (e.g., cognitive test)
- Sensitivity, specificity, accuracy, dice similarity index, false negative ratio (FNR) and false positive ratio (FPR) of the proposed segmentation method
- Visual rating scale used for validating the segmentation method (if any)

Extracted data were tabulated, synthesized, and evaluated for methodological flaws and applicability of the proposed techniques.

3. Results

3.1. Search results

The search yielded 2327 papers after removing 104 duplicate citations. We schematically represent the selection process in Fig. 2; we conducted it according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).

3.1.1. Exclusions

We removed 2268 papers after screening titles and abstracts, leaving 59 for full text screening. We excluded a further 19 for these reasons: validated the method using datasets with brain tumours (2); or MS patients only (6); method for segmenting lacunes (1); or perivascular spaces (3); or only small T2 hyperintensities (1); full text unavailable (1); sample size less than 20 (1); presented a tool for displaying but not segmenting WMH (1); modelling WMH distribution (1 study); or only quantifying longitudinal change (1). Also, three studies did not propose a new segmentation method of WMHs but compared the performance of existing machine learning based segmentation methods of WMHs (Dudar et al., 2017b; Kuijf et al., 2019; Rachmadi et al., 2017), leaving 37 studies for full analysis.

3.2. Studies characteristics

3.2.1. Individuals analysed

The 37 studies included analysed imaging data from approximately
7000 individuals. The exact number and quantitative breakdown of sample characteristics was not possible to ascertain because several studies used overlapping publicly available datasets lacking accompanying clinical or demographic metadata. One large-scale multicentre study involved 2781 subjects from 12 different sites (Schirmer et al., 2019); two studies analysed more than 500 individuals (Griffanti et al., 2016 (n = 559); Jiang et al., 2018 (n = 566)); eight studies used small relevant samples (i.e., fewer than 30 individuals with WMH of presumed vascular origin) (Ding et al., 2020; Rachmadi et al., 2020; Rincón et al., 2017; Roy et al., 2015; Stone et al., 2016; Van Opbroek et al., 2015a, b; Valverde et al., 2017); three additionally used data from MS patients (Rachmadi et al., 2020; Van Opbroek et al., 2015a, b).

3.2.2. Validity of the data extracted
Fourteen studies validated their results in data provided by Medical Image Computing and Computer Assisted Intervention (MICCAI) segmentation challenges (Roy et al., 2015; Sudre et al., 2015; Van Opbroek et al., 2015a, b; Wang et al., 2015; Valverde et al., 2017; Zhan et al., 2017; Knight et al., 2018; Li et al., 2018; Manjón et al., 2018; Moeskops et al., 2018; Sundaresan et al., 2019; Wu et al., 2019b; Liu et al., 2020). Nine of them also used additional datasets or patient data from clinics. Out of 37, twelve studies reported using data from prospective studies or clinics (Atlason et al., 2019; Bowles et al., 2017; Guerrero et al., 2017; Hong et al., 2020; Moeskops et al., 2018; Ling et al., 2018; Park et al., 2018; Qin et al., 2018; Rincón et al., 2017; Roy et al., 2015; Sundaresan et al., 2019; Wang et al., 2015). Four studies used data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (Sudre et al., 2017; Dadar et al., 2017b; Dadar et al., 2017a; Rachmadi et al., 2018, 2020). Two studies reported having selected the cases randomly (Atlason et al., 2019). One study stated that the data selection and manipulation were blinded to clinical information (i.e., avoided clinical review bias) (Dadar et al., 2017a). One study reported using data acquired at both 1.5 T and 3 T (see Table 2).

Data inclusion and exclusion criteria were not explained in 22/37 studies. Of the studies that reported demographic information, five recruited healthy controls (Griffanti et al., 2016; Sundaresan et al., 2019; Damangir et al., 2017; Rincón et al., 2017; Ding et al., 2020). One study stated that the data selection and manipulation were blinded to clinical information (i.e., avoided clinical review bias) (Dadar et al., 2017a). One study reported using data acquired at both 1.5 T and 3 T (see Table 2).

The magnet strength of the scanner used to acquire the data was reported in 35/37 studies (see Table 2). Twelve studies used data only acquired at 1.5 T, and twelve used data only acquired at 3 T. 11/37 studies used data acquired at both 1.5 T and 3 T (see Table 2).

We observed differential verification bias in 37 studies. These studies used different reference standards to verify segmentation methods’ performances; i.e., more than one reviewer was involved in manual WMH delineation of different datasets, or each dataset was delineated by a different person using different strategies, without stating the degree of inter-observer reliability or whether or not the final reference segmentation was agreed between the observers involved. Only 8/37
Table 2
Sample characteristics, type of segmentation algorithm proposed and average spatial agreement with reference segmentations in the relevant dataset used, from the 37 studies reviewed. Studies appear in alphabetical order of the first author’s surname by year of publication.

| STUDY / CODE REPOSITORY | SAMPLE SIZE | SAMPLE CHARACTERISTICS | TYPE OF SEGMENTATION | AVERAGE SPATIAL AGREEMENT WITH REFERENCE SEGMENTATION |
|-------------------------|-------------|------------------------|----------------------|-----------------------------------------------------|
| Roy et al. (2015)       | 24          | Selected from a larger cohort of elderly subjects with hypertension based on the burden of WMH. Further evaluation in data from 20 MS patients from the MICCAI MS Segmentation Challenge II 2008 (https://www.nitrc.org/projects/msseg) | Supervised | DSC 0.60–0.76 In MS patients: DSC 0.42, TPR 0.55, PPV 0.39 |
| Sudre et al. (2015)     | Not especi-fied | Simulated data from the BrainWeb project (Kwan et al., 1999), MS patient data from https://www.nitric.org/projects/msseg and data with age-related WMH from the MICCAI BrainS challenge | Unsupervi-sed | DSC 0.46, PPR 0.1, TPR 0.38, FNR 0.62 |
| Van Opbroek et al. (2015a) | 40          | 20 healthy elderly subjects from the Rotterdam scan study and 20 MS patients from https://www.nitrc.org/projects/msseg | Supervised | Total accuracy score 78 %, TPR 42.3 & 52.7 %, FPR 73.2 & 70.0 % |
| Van Opbroek et al. (2015b) | 40          | 20 healthy elderly subjects from the Rotterdam scan study and 20 MS patients from https://www.nitrc.org/projects/msseg | Supervised | – |
| Wang et al. (2015)      | 60 +        | Elderly patients scanned as part of normal patient care, aged 61–86 years (mean age 68.2 years), with various degrees of vascular white matter abnormalities. Further evaluation used data from 10 MS patients from the MICCAI MS Segmentation Challenge II 2008 (https://www.nitrc.org/projects/msseg) | Unsupervi-sed | DSC 0.81–0.84, FPR 0.13–0.24, FNR 0.14–0.2 In MS patients: TPR 0.38–0.40, PPV 0.36–0.48 |
| Zhan et al. (2015)      | 40          | Patient selection process not provided, mean age 62.2 ± 5.9 years | Unsupervi-sed | DSC 0.75, TPF 0.72, FPF 0.46 |
| (Damangir et al., 2017) | 119         | Patients from Kings Health Partners-Dementia Case Register with AD, MCI, and healthy controls, aged 76.4 ± 7.4 years, 56 % Females | Unsupervi-sed | DSC 0.85 – 0.91 |
| Griffanti et al. (2016) | 559         | Selection criteria not provided. Dataset 1 (neurodegenerative cohort)- 85 patients Dataset 2 (vascular cohort)-MRI data from 474 consecutive patients | Unsupervi-sed | DSC 0.76, FNR 0.25, FPR 0.22 |
| Stone et al. (2016)     | 24          | Traumatic brain injured patients aged 28–58 years (mean age 39.6 ± 8.1 years) with isolated traumatic lesions in the cerebral white matter, recruited based on potential concussive events | Supervised | Sensitivity 0.68, PPV 0.51 |
| Bowles et al. (2017)    | 127         | Selection criteria not provided. Heterogenous dataset containing imaging data from three different acquisition protocols | Unsupervi-sed | DSC 0.70 |
| Dadar et al. (2017a)    | 130         | Dataset 1: 80 subjects aged 70–90 years old with either normal cognition, MCI or AD. Dataset 2: 40 cognitively normal subjects at risk of AD aged 55–75 years Dataset 3: 10 subjects from ADNI2/GO selected to have different WMH burden Subjects for the RUN DMC study aged between 50 and 85 years with cerebral SVD on neuroimaging (appearance of WMH and/or lacunes) | Supervised | DSC 0.51–0.62 |
| Ghaffoorian et al. (2017) | 420       | Subjects for the RUN DMC study aged between 50 and 85 years with cerebral SVD on neuroimaging (appearance of WMH and/or lacunes) | Deep learning | DSC 0.78–0.79 |
| Rincón et al. (2017)    | 28          | 13 patients with cortical and lacunar ischemic infarctions, between 40 and 79 years of age, mini-mental state examination (MMSE) scores ≥ 23, no severe problems of language and visual/auditory neglect. 15 MCI patients meeting Petersen criteria. | Supervised | DSC 0.64 |
| Sudre et al. (2017)     | 85 +        | Two datasets of unspecified size: one from ADNI with minimal to none WMH load and other with a range of WMH burden. Additional Synthetic data derived from a non-specified number of data was used. | Unsupervi-sed | DSC 0.2 – 0.7 |
| Valverde et al. (2017)  | 20          | MRBrainS public database, selected with varying atrophy degrees and WMH loads | Unsupervi-sed | TPF 0.7–0.8 |
| Zhan et al. (2017)      | 104         | Dataset 1: 50 subjects from ACCORD-MIND mean age 62 years Dataset2: 54 image datasets from https://www.nitrc.org/projects/msseg | Supervised | DSC 0.76, TPR 0.83 |
| Diniz et al. (2018)     | 91          | Image data from a database with some of the volumes containing less than 10 lesions clusters while other volumes contains more than 100. | Deep learning | Sensitivity 78.79 %, Specificity 98.77 % |

(continued on next page)
| STUDY / CODE REPOSITORY | SAMPLE SIZE | SAMPLE CHARACTERISTICS | SEQUEN-CES AND FIELD STRENGTH | TYPE OF SEGMENTATION | AVERAGE SPATIAL AGREEMENT WITH REFERENCE SEGMENTATION |
|-------------------------|-------------|------------------------|-------------------------------|----------------------|--------------------------------------------------------|
| (Guerrero et al., 2017) | 167         | Patients with their first clinically evident non-disabling lacunar or mild cortical ischemic stroke | T1W, FLAIR / 1.5 T         | Deep learning       | DSC 0.607 |
| Jiang et al. (2018)     | 566         | Data from two sources: 1) Sydney Memory and Ageing Study (n = 166 aged 70–90 years old provided longitudinal data in 3 time points), 2) Older Australian Twins Study (n = 400 aged 65 and above). | T1W, FLAIR / 1.5 T & 3 T | Supervised          | DSC 0.85, Sensitivity 0.91, Specificity 0.99 |
| Knight et al. (2018)    | 96          | From 7 different sources, provided by different WMH and MS lesion segmentation challenges (Stynier et al., 2007) | FLAIR / 1.5 T & 3 T       | Supervised          | DSC 0.41–0.70 |
| Li et al. (2018)        | 170         | WMH segmentation challenge dataset (https://wmh.isi.uu.nl/). | T1W, FLAIR / 1.5 T & 3 T | Deep learning       | DSC 0.8, Recall 0.84 |
| Ling et al. (2018)      | 156         | CADASIL patients 18 years old and above. Mutation of gene NOTCH3 confirmed. 90 patients with 2D FLAIR images were aged 24–74 years, mean age 49 ± 11 years, F:M = 54:36. 66 patients with 3D FLAIR images were aged 35–81 years, mean age 57 ± 11 years, F:M = 46:20 | T1W, FLAIR / 1.5 T & 3 T | Supervised          | DSC 0.79 for 2D FLAIR images and 0.76 for 3D FLAIR images |
| Manjón et al. (2018)    | 128 +       | 128 subjects with a wide range of WMH load aged 38.6–92.1 years (M:F = 60:68) from Australian Imaging Biomarkers and Lifestyle (AIBL) study, and unspecified data from https://www.nitrc.org/project/rs/nnseg | T1W, FLAIR / 1.5 T & 3 T | Deep learning       | DSC 0.78 |
| Moeskops et al. (2018)  | 226         | Patient data from MRBrainS13 challenge and two other datasets: patients with type 2 diabetes mellitus, healthy controls and patients from memory clinic | T1W, FLAIR / 3 T          | Deep learning       | DSC 0.67 |
| Park et al. (2018)      | 148         | Diagnosis of migraine confirmed by two headache specialists mean age 44.4 (SD 12.4) years | T1W, FLAIR / 3 T          | Supervised          | TPR 0.7–0.91 |
| Qin et al. (2018)       | 88 (uses WM lesion 3D atlas from 277 FLAIR images) | Adults from mid to older ages with WMH without confounding radiological evidence of recent or old strokes | T1W, FLAIR / 1.5 T        | Supervised          | DSC 0.65–0.67, Precision 0.68, Recall 0.68–0.69 |
| Rachmadi et al. (2018)  | 288         | ADNI subjects randomly selected. From the subsample with ground truth segmentations (n = 20) 3 were cognitively normal, 12 had early MCI and 5 had late MCI | T1W, T2W, FLAIR / 3 T      | Deep learning       | DSC 0.54 |
| Atlason et al. (2019)   | 170         | AGES-Reykjavik study. Age 66–93 years at first visit, and WMH segmentation challenge dataset (https://wmh.isi.uu.nl/). | T1W, T2W, FLAIR / 1.5 T & 3 T | Unsupervised        | DSC 0.77, TPR 0.64 in AGES-Reykjavik dataset. DSC 0.53–0.67, TPR 0.25–0.40 in WMH segmentation challenge dataset (Provides ICC = 0.84, Pearson r = 0.86 with p < 0.001 (validation only used 144 images)) |
| Schirmer et al. (2019)  | 2783 (WMH extracted in 2533) | Acute ischaemic stroke patients ages 63.28 (SD 14.70) years, 61 % male, 10.6 % had a prior stroke | FLAIR / Field strength not specified | Deep learning       | DSC 0.77 |
| Sundaresan et al. (2019)| 133         | Five datasets: 1) Neurodegenerative cohort (n = 21, age range 65–86 years, mean age 77.1 ± 5.8 years F:M = 10:11), 2) Patients that recently experienced a minor non-disabling stroke or transient ischemic attack (n = 18, age range 50–91 years, mean age 73.27 ± 12.32 years, F:M = 7:11) 3) CADASIL patients (n = 15, age range 33–70 years, mean age 53.73 ± 11.31 years, F:M = 11:4) 4) Healthy controls (n = 19, age range 29–70 years, mean age 54.58 ± 11.25 years, F:M = 6:13) 5) MICCAI WMH segmentation challenge dataset (n = 60, selected from 3 different primary cohorts, https://wmh.isi.uu.nl/). | T1W, T2W, FLAIR / 1.5 T & 3 T | Supervised          | TPR 0.73–0.98 |
| Wu et al. (2019)        | 135         | Data from BIOCARD study. Subjects either cognitively normal (n = 113) or MCI (n = 22), based on the (NIA/AA) research diagnostic criteria | T1W, FLAIR / 3 T          | Supervised          | DSC 0.62, FPR 0.35, FNR 0.37 |
| Wu et al. (2019b)       | 60          | Data from a MRI database. Subjects included patients with a history of stroke, TIA, or both. The database contains data from the MICCAI WMH segmentation challenge (https://wmh.isi.uu.nl/). Demographic data not provided | T1W, FLAIR / 1.5 T & 3 T  | Deep learning       | DSC 0.78, Recall 0.81 |
| Ding et al. (2020)      | 20          | Participants from ongoing normal ageing study. Average age 81.2 (SD 7.15) years. 70 % Females. 85 % white and 15 % African-Americans | T1W, FLAIR / 3 T          | Supervised          | DSC 0.78, FPR 0.009, PPV 0.85, TPR 0.70 |

(continued on next page)
information about pre-processing steps performed before segmentation brain extraction, intensity inhomogeneity correction, noise reduction. These were often registration, known as sequences. 28 used a combination of more than one sequence (i.e., also Rachmadi et al., 2018, 2020; Valverde et al., 2017). One study (Ling et al., 2020) evaluated different configurations of the method described by Griffanti et al. (2016) making recommendations of its use. We present risk of bias assessment of the 37 included studies using QUADAS 2 tool in Table 3. Out of the 37 studies, only 7 were judged as having low risk of bias overall.

3.4. Pre-processing methods

All studies which reported ground truth generation details, validated the WMH segmentation method with ground truth binary masks, generated using the FLAIR MRI sequence. However, only three studies reported having used only the FLAIR sequence in their segmentation framework (Diniz et al., 2018; Knight et al., 2018; Schirmer et al., 2019). Of the rest (i.e., 34/37) which described using data from different sequences, 28 used a combination of more than one sequence (i.e., also known as “multispectral approach”), generally T1-weighted and FLAIR, to generate the final outcome. In general, after MRI acquisition, various pre-processing steps were conducted. These were often registration, brain extraction, intensity inhomogeneity correction, noise reduction and intensity normalisation. Table 4 summarises the publicly available tools used in the studies’ pipelines and Table 5 summarises the pre-processing steps used by each study. Only one study reported having conducted all the above-mentioned pre-processing steps prior to the segmentation method (Manjón et al., 2018), and one did not provide any information about pre-processing steps performed before segmentation (Liu et al., 2020). The latter selected MRI slices from already brain-extracted images downloaded from an image data repository, without specifying how the slice selection was performed (i.e., by visual inspection or automatically). Slice selection excluded 81 slices with haemorrhagic stroke and those at the top and bottom of the brain, which are more prone to have confounding artefacts.

Most studies included in the review used linear transformation models to co-register the different image sequences from each subject data (Fig. 4 (left hand side), Table 5). From the 27/37 studies that provided information on the tool used for image registration, ten used FSL FLIRT (Jenkinson et al., 2002) (Qin et al., 2018; Griffanti et al., 2016; Guerrero et al., 2017; Sundaresan et al., 2019; Damangir et al., 2017; Rachmadi et al., 2018, 2020; Valverde et al., 2017), one study used elastix (Shamonin et al., 2014) (Moeskops et al., 2016), five used SPM (SPM - Documentation, 2020) (Wu et al., 2019a; Roy et al., 2015; Jiang et al., 2018; Knight et al., 2018; Valverde et al., 2017), one study used elastix (Shamonin et al., 2014) (Moeskops et al., 2018), one study used 3DSlicer (Documentation/4.10/Training - Slicer Wiki, 2020) (Rincon et al., 2017), and one study used MIRTK (MIRTK – BioMedIA, 2020) (Bowles et al., 2017). The skull stripping method for removal of non-brain tissues was reported in 17/37 studies: FSL BET (Jenkinson et al., 2004), ANTs Advanced Normalization Tools (ANTs), 2020 (Schirmer et al., 2019; Manjón et al., 2018; Stone et al., 2016), five used SPM (SPM - Documentation, 2020) (Wu et al., 2019a; Roy et al., 2015; Jiang et al., 2018; Knight et al., 2018; Valverde et al., 2017), one study used elastix (Shamonin et al., 2014) (Moeskops et al., 2018), one study used 3DSlicer (Documentation/4.10/Training - Slicer Wiki, 2020) (Rincon et al., 2017), and one study used MIRTK (MIRTK – BioMedIA, 2020) (Bowles et al., 2017). The skull stripping method for removal of non-brain tissues was reported in 17/37 studies: FSL BET (Jenkinson et al., 2004), ANTs Advanced Normalization Tools (ANTs), 2020, OptiBET (Lutkenhoff et al., 2014), Neuron BE, SPM, Freesurfer, MONT STR and MIRcroVisi bleHuman, 2021). Intensity inhomogeneity correction was reported in 17/37 studies, and it was always performed using a well-known tool: N3 (or N4), SPM, FSL FAST (Zhang et al., 2001) or the Nu estimate. N3 (Sied et al., 1998) and its newer version N4
### Table 3
Risk of Bias Assessment of the studies reviewed using QUADAS 2 Tool ([https://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/quadas2.pdf](https://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/quadas2.pdf)). Studies appear in alphabetical order of the first author’s surname by year of publication.

| STUDY                        | RISK OF BIAS | APPLICABILITY |
|------------------------------|--------------|---------------|
|                              | PATIENT SELECTION | INDEX TEST | REFERENCE TEST | FLOW & TIMING | PATIENT SELECTION | INDEX TEST | REFERENCE TEST |
| Roy et al. (2015)            | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Sudre et al. (2015)          | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Van Opbroek et al. (2015a)   | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Van Opbroek et al. (2015b)   | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Wang et al. (2015)           | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Zhan et al. (2015)           | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| (Damangir et al., 2017)      | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Griffanti et al. (2016)      | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Stone et al. (2016)          | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Bowles et al. (2017)         | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Dadar et al. (2017a)         | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Ghafoorian et al. (2017)     | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Rincón et al. (2017)         | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Sudre et al. (2017)          | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Valverde et al. (2017)       | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Zhan et al. (2017)           | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Diniz et al. (2018)          | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| (Guerrero et al., 2017)      | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Jiang et al. (2018)          | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Knight et al. (2018)         | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Li et al. (2018)             | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Ling et al. (2018)           | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Manjón et al. (2018)         | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |
| Moeskops et al. (2016)       | ☒ ☒ ☒ ☒ ☒ | ☒ ☒ ☒ ☒ ☒ | ☒ | ☒ | ☒ | ☒ |

(continued on next page)
(Tustison, 2010), were the tools most commonly used for intensity inhomogeneity correction (Stone et al., 2016; Wu et al., 2019a; Bowles et al., 2017; Dadar et al., 2017a; Van Opbroek et al., 2015a, b; Damangir et al., 2017; Roy et al., 2015; Wang et al., 2015; Zhan et al., 2015, 2017; Atlason et al., 2019; Ding et al., 2020). Non-local means (Coupe et al., 2008) was the only filtering technique used by the two studies that reported having included noise removal within their pre-processing steps (Manjon et al., 2018; Dadar et al., 2017a). Neither of these two studies selectively applied the filtering after analysing the signal. The 23/37 studies that provided information on intensity normalisation, reported the use of either variance / linear scaling or histogram matching, with variations in their implementation.

3.5. Segmentation methods

From the studies included, 73 % (i.e., 27/37) proposed a supervised segmentation method. Of them, 37 % (i.e., 10/27) used deep learning. From the 10 studies that proposed an unsupervised segmentation method (i.e., 27 % of the total number of studies included), one used deep learning (Atlason et al., 2019). In total, eleven studies used Convolutional Neural Networks (Rachmadi et al., 2018; Li et al., 2018; Guerrero et al., 2017; Moeskops et al., 2018; Ghafoorian et al., 2017; Hong et al., 2020; Manjon et al., 2018; Wu et al., 2019a; Liu et al., 2020; Diniz et al., 2018; Schirmer et al., 2019), four studies proposed a method based on k-nearest neighbours (k-NN) (Sundaresan et al., 2019; Jiang et al., 2018; Griffanti et al., 2016), four studies proposed regression models (Knight et al., 2018; Dadar et al., 2017a; Zhan et al., 2017; Ding et al., 2020), and three studies used Random forest (RF) in their proposed algorithms (Stone et al., 2016; Park et al., 2018; Roy et al., 2015). Two studies proposed a method based on Fuzzy C mean algorithm (Zhan et al., 2015; Valverde et al., 2017) and three proposed improvements to a Gaussian Mixture Model framework (Sudre et al., 2015, 2017; Fiford et al., 2020), both unsupervised methods. We summarize the segmentation methods included in the reviewed studies in Fig. 4 (right hand side).

### 3.5.1. Supervised WMH segmentation methods

#### 3.5.1.1. k-Nearest neighbours (k-NN)

k-NN is a well-established pattern recognition method that, for WMH segmentation, compares each voxel’s spatial (i.e., location) and intensity features with those extracted from a training set, and assigns a probability of being (or not) WMH based on the result. This algorithm was first proposed for this task in 2000 (Warfield et al., 2000), further evaluated in 2004 (Anbeek et al., 2004) and improved by additionally using spatial tissue type priors in further works (De Boer et al., 2007; Steenwijk et al., 2013). Three of the four papers included in this review that use this method (Ling et al., 2018; Griffanti et al., 2016; Sundaresan et al., 2019), use the implementation Brain Intensity Abnormality Classification Algorithm (BIANCA) of the FMRIB Software Library (FSL). BIANCA (Griffanti et al., 2016) is a versatile, easy to use, freely available implementation, which offers

---

**Table 3 (continued)**

| STUDY                  | PATIENT SELECTION | INDEX TEST | REFERENCE TEST | FLOW & TIMING | APPLICABILITY |
|------------------------|-------------------|------------|----------------|--------------|---------------|
|                        | Could the selection of patients have introduced bias? | Could the method have introduced bias? | Is the ref. std. likely to be correct? | Is the ref. std. mani-pulated blind to the index test? | Did all data had the same ref. std.? | Were all patients included? | The included patients match the review question? | Are there concerns re. applicability? | Are there concerns re. reproducibility? |
| Park et al. (2018)     | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Qin et al. (2018)      | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Rachmadi et al. (2018) | 😈          | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Atlason et al. (2019)  | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Schirmer et al. (2019) | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Sundaresan et al (2019)| 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Wu et al. (2019a)      | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Wu et al. (2019b)      | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Ding et al. (2020)     | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Fiford et al. (2020)   | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Hong et al. (2020)     | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Liu et al. (2020)      | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |
| Rachmadi et al. (2020) | 😈            | 😈          | 😈              | 😈        | 😈            | 😈          | 😈              | 😈            | 😈       |

- High 😈 - Unclear 😈 - Low 😈
Table 4
Publicly available software resources and tools used in the selected studies.

| Software library / Repository | Tool                                         | Use                                                                 | References / Documentation |
|-------------------------------|----------------------------------------------|----------------------------------------------------------------------|--------------------------|
| FMRIB software library (FSL)  | Brain Extraction Tool (BET)                  | Removal of non-brain tissues                                         | Smith (2002); Jenkinson et al. (2004) |
|                              | FMRIB’s Linear Image Registration Tool (FLIRT) | Linear (affine) intra- and inter-modal brain image registration   | Jenkinson and Smith (2001); Jenkinson et al. (2002) |
|                              | FMRIB’s Automated Segmentation Tool (FAST)   | Tissue segmentation and correction for bias field inhomogeneities  | Zhang, Brady & Smith (2001) |
|                              | Set of MATLAB scripts and functions integrated in a graphic interface environment that can be used independently |                                                                     | (SPM - Documentation, 2020) |
| Statistic Parametric Mapping (SPM) | antiRegistration                          | Part of the ANTs suite of image registration tools                  | (Advanced Normalization Tools (ANTS) - SourceForge.net, 2020) |
|                              | N3, N4                                      | Correction for bias field inhomogeneities                            | (ANTSx/ANTS, 2020) Advanced Normalization Tools (ANTS), 2020 |
|                              | bricBET (MATLAB wrappers to FSL-BET)        | Generation of the intracranial volume binary mask from combinations of multiple MRI sequences | https://sourceforge.net/projects/bric1936/files/Documentation/1936/files/Documentation/2020 |
|                              | Modules for Image registration               | Image registration                                                   | (Documentation/4.10/Training - Slicer Wiki, 2020) |
|                              |                                             |                                                                      |                          |
| 3DSlicer                      |                                             |                                                                      |                          |
| Medical Image Registration Tool (MIRTK) |                                             |                                                                      |                          |
| OptiBET                       |                                             |                                                                      |                          |
| Elastix                       |                                             |                                                                      |                          |
| Niftyreg                      |                                             |                                                                      |                          |
| MRIcro                       |                                             |                                                                      |                          |
| FreeSurfer                    |                                             |                                                                      |                          |
| Multi-CONtrast brain STRipping (MONSTR) |                                             |                                                                      |                          |
| Analysis of Functional Neuroimages (AFNI) |                                             |                                                                      |                          |

3.5.1.2. Large margin classifiers. Large margin algorithms maximise the margin around the decision boundary of a classifier to reduce the uncertainty in the classification, handling well, high-dimensional data. Qin et al. (2018) developed a supervised large margin algorithm (SLM) followed by a semi-supervised large margin algorithm (SSLM) in a framework that modifies a self-guided labelling procedure, namely unsupervised one-class learning (UOCL) (Liu et al., 2014), which discovers potential “outliers” in the data, being the WMH. Qin et al. (2018) introduced a new term in the objective function of the UOCL that maximises the average margin between the hyperintensities (i.e. considered outliers) and the decision boundary. The general SLM classifier minimises the objective function using a conjugate gradient method to learn from the training set and provides a rough WMH segmentation map. The SSLM, then, refines the given labels on the target data.

3.5.1.3. Multi-atlas segmentation. Wu et al. (2019a) presented a framework that simultaneously segments the brain and detects WMH. The proposed multi atlas-based detection and localization (MADL) framework uses a multi-atlas likelihood fusion approach to segment the brain tissues and structures, and identify WMH. It uses a multi-atlas library generated from 15 FLAIR images with minimal WMH load and atrophy ranging from minimal to moderate. The Bayes maximum a posteriori estimation generates a maximum posterior probability value for each voxel, of belonging to a certain (atlas) label. The WMH are identified as voxels with maximum posterior probability values below certain threshold empirically determined.
Table 5
Pre-processing in the studies involved. Studies appear in alphabetical order of the first author’s surname by year of publication (as per previous tables).

| STUDY                  | REGISTRATION (details reported)                                   | BRAIN EXTRACTION                  | INTENSITY INHOMOGENEITIES CORRECTION | NOISE REDUCTION                  | INTENSITY NORMALISATION                  |
|------------------------|-------------------------------------------------------------------|-----------------------------------|--------------------------------------|-----------------------------------|------------------------------------------|
| Roy et al. (2015)      | Rigid-body – uses SPM 8                                          | FSL-BET                           | N3                                   | Information not provided          | Histogram matching (Nylil and Udupa, 1999) |
| Sudre et al. (2015)    | Intra-subject inter-modality co-registration, and statistical atlases warped to observed data using niftyreg. | Performed using STEPS (Cardoux et al., 2013) followed by non-brain tissue mask filling | Information not provided          | Information not provided          | Intensity rescaling from 0 to 1.          |
| Van Opbroek et al. (2015a) | Information not provided                                       | Information not provided          | N4                                   | Information not provided          | Three normalisation algorithms were evaluated: 1) Range-matching (maps the 4th and the 96th percentage of intensity within the brain mask to 0 and 1. 3) Method 1 followed by mapping of every tenth percentile within 0 and 1 to the mean intensity over all (training and target) images. |
| Van Opbroek et al. (2015b) | Information not provided                                       | FSL-BET                           | N4                                   | Information not provided          | Range-matching procedure that scaled the voxels within a mask such that the voxels between the 4th and 96th percentage in intensity are mapped between 0 and 1 | |
| Wang et al. (2015)     | Information not provided                                         | Brain Extraction Tool in MRicro   | N3                                   | Information not provided          | Image intensity rescaling from 0 to 255                                      |
| Zhan et al. (2015)     | A mutual information-based registration method (Viola and Wells, 1997) | FSL-BET                           | N3                                   | Information not provided          | Information not provided          |
| Damangir et al., 2017  | Intra-subject, rigid, 3-D with mutual information - uses FSL FLIRT | FSL-BET                           | N3                                   | Information not provided          | Information not provided          |
| Griffanti et al. (2016) | Linear, with trilinear interpolation (in multisequence evaluation using T1-weighted and FLAIR) - uses FSL-FLIRT. | Information not provided          | Information not provided            | Information not provided          | Used variance scaling                  |
| Stone et al. (2016)    | Rigid-body, linear intra-subject, of FLAIR, T1- and T2-weighted sequences – uses ANTs | ANTs                              | N4 bias correction                   | Information not provided          | Normalized to the intensity range 0–1.                                      |
| Bowles et al. (2017)   | Rigid-body, linear, of T1-weighted to FLAIR followed by free-form deformation between T1-weighted images in FLAIR image space and an MNI template –uses MIRTK suite. | pinpan algorithm (uses label propagation and group agreement) (Heckemann et al., 2015) | N4 algorithm                       | Information not provided          | Linear scaling using the mean intensity of the healthy (i.e., gray and white matter) tissue in the cohort (estimated having the value of 1000). (Huppertz et al., 2011) |
| Dadar et al. (2017a)   | Intra-subject, linear rigid body, of T2-weighted, PD and FLAIR to T1-weighted, then non-linear based on intensity correlation coefficient to ADNI 150 template | Information not provided          | N3                                   | Automatic multi-threaded denoising method based on non-local means filtering (Manjon et al., 2010) | Normalized per patient to be within the range of [0, 1] |
| Ghafoorian et al. (2017) | Rigid-body, linear, of T1-weighted to FLAIR with trilinear interpolation and mutual information optimization – uses FSL-FLIRT | FSL-BET on T1-weighted images     | FSL-FAST                            | Information not provided          | Normal white matter modelled for Gaussian fit to generate parameters for FLAIR intensity normalisation using the EzyFit Toolbox (EzyFit 2.44, 2020) |
| Rincón et al. (2017)   | Linear, intra-subject, of T1-weighted to FLAIR – uses BRAINSFit from 3D-Slicer | White matter tissue and ventricles segmented using Freesurfer | SPM 8                               | Information not provided          | Normal white matter modelled for Gaussian fit to generate parameters for FLAIR intensity normalisation using the EzyFit Toolbox (EzyFit 2.44, 2020) |
| Sudre et al. (2017)    | Affine, linear, intra-subject, of FLAIR to T1-weighted. Non-linear (niftyreg) used for generating synthetic data. | Performed, but information not provided | Performed, but information not provided | Information not provided          | Information not provided          |
| Valverde et al. (2017) | Affine followed by non-linear registration of the MNI-ICBM152 brain template to the native T1-weighted space - used niftyreg followed by the registration tool in SPM12. | Information not provided          | Information not provided            | Information not provided          | Information not provided          |
| Zhan et al. (2017)     | Information not provided                                         | FSL-BET on T1-weighted images     | N3 bias field correction             | Information not provided          | Linear contrast adjustment to match the intensities of the training and testing datasets |
| Bandeira Diniz et al. (2018) | Not applicable (uses only FLAIR images) | Algorithm by Bauer et al. (2011)  | Information not provided            | Information not provided          | Histogram matching algorithm      |

(continued on next page)
Table 5 (continued)

| STUDY | REGISTRATION (details reported) | BRAIN EXTRACTION | INTENSITY INHOMOGENEITIES CORRECTION | NOISE REDUCTION | INTENSITY NORMALISATION |
|-------|---------------------------------|------------------|--------------------------------------|----------------|-------------------------|
| (Guerrero et al., 2017) | T1-weighted registered to FLAIR - uses FSL-FLIRT | Information not provided | Information not provided | Information not provided | Tissue intensities normalised to have zero mean and standard deviation of one (excluding the background). Information not provided |
| Jiang et al. (2018) | Rigid body, linear, of FLAIR to T1-weighted and from T1-weighted space to the Diffeomorphic Anatomical Registration Through Exponentiated (DARTEL) space - uses SPM 12 | SPM 12 - used brain mask in DARTEL space to remove non-brain tissues | FSL-FAST | Information not provided | Information not provided |
| Knight et al. (2018) | Training images are resampled (trilinear) to 1.5 mm isotropic voxel resolution in the MNI-ICBM152 space – uses the Segment tool in SPM 12 | Segment tool in SPM 12 | Segment tool in SPM 12 | Information not provided | Information not provided |
| Li et al. (2018) | (Retrospective data, intra-subject MRI sequences previously co-registered) | (data was already brain extracted) | SPM12 | Information not provided | Gaussian intensity normalisation |
| Ling et al. (2018) | Linear, of FLAIR to T1-weighted space and from T1-weighted to MNI-ICBM152 standard space | FSL-BET | Information not provided | Information not provided | Variance scaling within FSL-BIANCA |
| Manjon et al. (2018) | Linear registration to the MNI-ICBM152 space – uses ANTs | FSL-BET Information not provided | Information not provided | Information not provided | All brain voxels intensities were divided by the median intensity within the brain region. Resulting intensities were squared to enhance image contrast |
| Moeskops et al. (2018) | Intra-subject, of T1-weighted (with and without inversion recovery pulse) to FLAIR – uses elastix (Klein et al., 2019) | FSL-BET | Information not provided | Information not provided | Information not provided |
| Park et al. (2018) | Rigid-body – uses FSL-FLIRT | Information not provided | Information not provided | Information not provided | Information not provided |
| Qin et al. (2018) | Rigid-body, linear, of T1-weighted and FLAIR - uses FSL-FLIRT | Information not provided | As per the unified segmentation algorithm in SPM5 | Information not provided | Information not provided |
| Rachmadi et al. (2018) | Rigid-body, linear - uses FSL-FLIRT | OptiBET | Information not provided | Information not provided | Histogram matching (Nyúl and Udupa, 1999) |
| Atlason et al. (2019) | Rigid registration to the MNI-ICBM152 template | skull removal using MONSTR inhomogeneity correction using N4 integrated into SegAE | Information not provided | Information not provided | Information not provided |
| Schirmer et al. (2019) | Affine – uses ANTs | (Neuron-BE) that used a 2D U-Net convolutional neural network architecture | Information not provided | Information not provided | Information not provided |
| Sundaresan et al. (2019) | Rigid-body, linear – uses FSL FLIRT | Information not provided | Information not provided | Information not provided | Information not provided |
| Wu et al. (2019a) | FLAIR to T1-weighted previously mapped in MNI-ICBM152 coordinates - uses SPM (version not specified) | N4 bias correction | Information not provided | Information not provided | Histogram matching |
| Wu et al. (2019b) | Information not provided (Retrospective data, intra-subject MRI sequences previously co-registered) | FSL-BET | Information not provided | Information not provided | Gaussian intensity normalization (z-scores) |
| Ding et al. (2020) | Linear within-subject co-registration of FLAIR and T1W sequences using FSL-FLIRT through fsl (Muchelli et al., 2019) | FSL-BET followed by erosion of the brain mask through fsl (Muchelli et al., 2019) | N4 | Information not provided | z-scores calculated over the entire brain-only intensity signal |
| Fiford et al., 2020 | Information not provided | Method proposed by Van Leemput et al (1999) | Information not provided | Information not provided | Information not provided |
| Hong et al. (2020) | Rigid-body linear within-subject co-registration of TIW and FLAIR sequences | Information not provided | Information not provided | Information not provided | Information not provided |
| Liu et al. (2020) | Information not provided | Information not provided | Information not provided | Information not provided | Information not provided |
| Rachmadi et al. (2020) | Rigid-body linear within-subject co-registration of TIW and FLAIR sequences using FSL-FLIRT | Information not provided | Information not provided | Information not provided | Information not provided |
3.5.1.4. Feature filters. Rincón et al. (2017) present an object-based segmentation framework, namely amorphous object segmentation in 2D (AMOS -2D). This method uses a multi-level information approach consisting of a hierarchical multi-threshold WMH segmentation followed by an object-based filter that reduces the number of false-positives. After pre-processing T1-weighted and FLAIR images, AMOS-2D applies white-matter Gaussian modelling to determine the intensity distribution of the WMH. An initial WMH mask is generated using multi-threshold segmentation, which combines single grey-scale thresholding with a seed-based thresholding. In the latter, the higher threshold (i.e. seed) acts as WMH detector and the lower threshold (i.e. region) refines the contours. The optimum thresholds are determined ad-hoc from the training dataset. The filter that refines the “initial” WMH mask is an object-based classifier that uses support vector machine (SVM). The feature vector for this classifier initially consisted of 178 features, including normalised intensity, others derived from applying connected-component analysis, distance to white matter contour, distance to white matter skeleton, distance to ventricles, among others not specified. The dimensionality of the initial feature vector was reduced using correlation-based feature selection.

Roy et al. (2015) present two filtering approaches: one for generating probabilistic regions of interest (i.e. weighted candidate voxels) for the segmentation algorithm to operate, and another to post-process the classifier’s output. The first are contrast-based global probabilistic maps generated from a feature set containing enhanced intensity, anatomical and spatial information, and the second is an edge potential function based Markov Random Field model, which is used to remove false positives and obtain the final output.

3.5.1.5. Regression models. Dadar et al. (2017a) proposed a multispectral linear regression classifier that uses the least-squares parameters estimation to segment WMH. It combines intensity and location features from FLAIR, T1-, T2- and PD-weighted MRI and manually labelled training data, to provide a continuous subject-specific WMH map displaying different levels of tissue damage along with a binary segmentation.

Knight et al. (2018) developed a supervised logistic regression framework exclusively for FLAIR sequences, called Voxel-Wise Logistic Regression. This method modifies the open source Lesion Segmentation Tool (LST) LPA by estimating the voxel-wise logistic regression parameters simultaneously across the image space for facilitating convergence during the parameters’ estimation, instead of randomly sampling the image space. The logistic model, trained using the standardised FLAIR intensity levels of a training set, generates a set of parameters that are subsequently smoothed for their use in the lesion prediction for new images.

Zhan et al. (2017) developed a supervised method that integrated the multi-sequence and spatial information in a Bayesian framework for WM lesion detection from multi sequence MR images. The proposed method is based on a three-step approach: 1) multinomial logistic regression is employed to learn the conditional probability distributions of WMH and brain tissues from training data; 2) spatial information from Markov random field priors is merged with multi sequence information in the Bayesian framework to improve the accuracy of WMH segmentation; and 3) pathology background information is used to reduce false positives.

Ding et al., 2020 present a supervised segmentation method called OASIS-AD. This approach is derived from a previous scheme (i.e., OASIS, Sweeney et al., 2013) developed for MS lesion segmentation, which uses a logistic regression model involving several imaging modalities to determine the probability of a voxel being WMH or not. This model uses as input brain-extracted and normalised image data. The enhanced version OASIS-AD additionally erodes the brain-extracted binary mask generated in the pre-processing step and refines the probability map obtained from the regression model by applying a nearest neighbour feature construction approach that uses FSL-FAST (Zhang et al., 2001), followed by a Gaussian filter.

3.5.1.6. Random Forest (RF). Park et al. (2018) present a machine learning based pipeline called DEWS (DEep White-matter hyper-intensity Segmentation framework). The authors segment the normal appearing white matter using FSL-FAST and use a combination of morphological operations and multi-level thresholding and inter-sequence registration to generate a normal white matter space that contained only deep WMH clusters in the FLAIR space. Then, a RF classifier uses size, texture and multi-parametric intensity statistical parameters from deep WMH (from a training set) as features for detecting small, superficially located deep WMH.

Stone et al. (2016) propose a multispectral framework that concatenates two RF classifiers, which the authors refer as a “two-stages” scheme. The first stage uses image intensity, symmetry, tissue segmentation voxel-wise probabilities, distance maps and neighbourhood statistics from the training data as features. These are used to produce the voxel-wise ‘voting maps’ (i.e. the classification count of each decision tree for each tissue label) of the first RF classifier for their use as tissue priors in a second multispectral 6-tissue segmentation that additionally
uses a Markov Random Field as spatial prior. The second stage uses all Stage 1 features plus the Stage 1 voting maps and the resulting posterior probability images as features for the second RF classifier. The whole framework is constructed on Advanced Normalization Tools (ANTs) and ANTsR toolkits. Stone et al. (2016) suggested that proposed supervised method is suitable for large dataset. However, this method is tested in a small sample size. 

Roy et al. (2015) use a set of nine features as input to the RF classifier. The first eight features contain multi-sequence (i.e. from T1-weighted and FLAIR) intensity, anatomical and spatial information per voxel. These are generated from probability maps of cerebrospinal fluid, grey and white matter, and normalised (x,y,z) coordinates in the MRI 152 space. The last feature is the global reference points-based contrast result from the filtering technique referred previously.

3.5.1.7. Support vector machine (SVM). Van Opbroek et al. (2015a, b) evaluate different transfer-learning approaches in linear and non-linear SVM classifiers, all consisting of different strategies for weighting the feature vector. Both studies use data from different datasets acquired under different transfer protocols and conclude that their transfer learning strategy (i.e. weighting the feature vector) outperforms the conventional SVM using non-weighted features. In Van Opbroek et al. (2015a) authors evaluate two feature sets: one of size 6 and other of size 33. The former uses the intensity and x,y,z voxel coordinates of cerebrospinal fluid, white and grey matter probabilistic segmentations in FLAIR and the latter uses the same features but also for T1- and T2/PD-weighted, using Gaussian kernels of $\sigma = 0.5, 1$ and 2 mm$^3$. In Van Opbroek et al. (2015b), the authors add the gradient magnitude and the Laplacian of the normalized intensities after convolution with the Gaussian kernel at different scales and recommend using always a feature vector higher than 10 in size. 

Van Opbroek et al. (2015a) assign weights to each feature of the feature vector in a way that the sum of all weights equals the total number of training samples, and combine training data with the same intensity distribution with data with different distribution in three weighting schemes: 1) Weighted SVM; 2) Re-weighted SVM; and 3) TransAdaBoost. In (1) lower weights are assigned to mismatched training data with different distribution. In (2) the misclassified lower weights (i.e. from (1)) are iteratively reduced. TransAdaBoost increases the weights of misclassified same-distribution data and reduces those from mismatched different-distribution data, but this scheme was the worst performer. In the same study authors also evaluate the namely “Adaptive SVM” that uses a weighted vector from same-distribution data for training and is tested from different-distribution data. 

Van Opbroek et al. (2015b) rather evaluates three different point distribution functions (PDFs) dissimilarity measures to generate the optimal weights for the Weighted SVM classifier – the winner scheme from those evaluated in (2015a)-, which in this case uses a Gaussian kernel. The weights are chosen in an unsupervised manner, by minimizing the difference between the PDFs of the weighted training images and the PDF of the target image. The three PDF dissimilarity measures evaluated are: 1) the Kullback-Leibler divergence; 2) the Bhattacharyya distance; and 3) the squared Euclidean distance. The optimal weights are determined by minimizing these three dissimilarity criteria while constraining them to the range [0,1] and that the norm of all the weights should be 1, using the interior-reflective Newton method (Coleman and Li, 1996).

3.5.1.8. Neural networks. Moeskops et al. (2018) evaluated the 3-pathway multi-scale (i.e. patch-wise) convolutional neural network (CNN) scheme developed by the same group in 2016 for segmenting normal tissues in neonatal and young adults (Moeskops et al., 2016) to segment WMH in addition to normal tissues in MRI scans for older individuals / patients. In this occasion, the scheme uses the T1-weighted, T2-weighted, FLAIR and T1-weighted inversion recovery (IR) images as input. Along with WMH, the scheme segmented normal-appearing white matter, cortical grey matter, basal ganglia, thalamus, cerebellum, brain stem, lateral ventricular cerebrospinal fluid, and peripheral cerebrospinal fluid.

Bandeira Diniz et al. (2018) use Simple Linear Iterative Clustering (SLIC) to group pixels based in their location and intensities and generate candidates to lesion / non-lesion regions in each FLAIR axial slice. Authors design a single-pathway CNN for extracting implicit features from the “superpixels” of the FLAIR axial slices presented as input and classify them in lesion regions or non-lesion regions. The CNN seems to have a linear deep architecture, developed ad-hoc for this purpose. This approach resulted efficient in heterogeneously sourced data, reporting a negligible number of false positives.

Rachmadi et al. (2018) proposed an adaptation of a dual-pathway CNN scheme developed for segmenting brain lesions with considerable mass effect (Ramittas et al., 2017) to segment WMH. The authors introduced a way to integrate spatial information to the CNN scheme for WMH segmentation called global spatial information (GSI), and evaluate the performance of two configurations (i.e. with 8 and 5 convolutional layers) using only FLAIR vs. using a combination of T1-weighted and FLAIR, and repeated the experiments using a single-pathway CNN architecture with and without GSI. Authors recommend the use of GSI in a multispectral (i.e. using more than one MRI sequence) dual-pathway scheme of the 2D CNN architecture evaluated.

Manjon et al. (2018) present an ensemble of patch-wise neural network classifiers for segmenting WMH on FLAIR images. After a lesion candidate ROI selection, a feature vector containing 58 features (voxel intensities from $3 \times 3 \times 3$ and $5 \times 5 \times 5$ pixels, 3 spatial coordinates and one a priori lesion probability) is used by an ensemble of two one-hidden layer feedforward multilayer perceptron which performs the classification. The study evaluates two ways of configuring this ensemble: bagging (Bootstrap aggregating) and boosting. The first approach averages the outputs of the two neural network classifiers, independently trained on different randomly selected datasets. The second approach uses the output from one classifier to improve the next one by either iteratively giving more weight in the next classifier, to the samples wrongly classified in the first one, or non-randomly selecting (i.e. on the training dataset) with higher probability samples wrongly classified previously.

Guerrero et al. (2017) use the UResNet CNN architecture to segment WMH and distinguish them from stroke lesions. This method comprised an analysis path that gradually learned low- and high-level features, followed by a synthesis path, that gradually combined and up samples the low and high-level features into a class likelihood semantic segmentation. The authors confirmed that the CNN architecture performed well compared to other state of the art algorithms.

Li et al. (2018) propose a method using a 19-layer deep fully CNN scheme. In this method, WMH detected based on convolution-deconvolution architecture with long-range connections which simultaneously classified each pixel and locates objects of an input image. The scheme used ensemble models with random parameter initialization and shuffled data for voting the pixel labels in the final evaluation, all which conferred good adaptability on multi-scanners and protocols and helped reduce overfitting. The authors pointed out that FLAIR and T1 sequences provide complementary information to detect WMH.

Schirmer et al. (2019) incorporate a deep learning CNN previously proposed by Dalca et al. in 2014 (Dalca, 2014), in a pipeline consisting of: 1) brain extraction using only clinical FLAIR images; 2) intensity normalisation to accommodate for multi-site heterogeneity; and 3) automatic atlas-based segmentation of WMH.

Ghafoorian et al. (2017) implement several deep CNN architectures which considered multi-scale patches or explicit location features while training, to integrate the anatomical location information into the network. The authors point out that the CNNs which incorporated location information significantly outperformed a conventional
segmentation method with hand-crafted features and CNNs that did not integrate location information.

Wu et al. (2019b) modified the U-Net CNN architecture by skipping connections between the down- and up-sampling convolutional branches of the original model, and named their model Skip Connection U-Net (i.e., SC U-Net). SC U-Net additionally connects the outputs of the 4th, 7th, 10th and 13th layers in the down-sampling convolutional branch of the original model to the outputs of the 15th, 18th, 21th and 24th layers in the up-sampling convolutional branch, and feeds them (i.e., the outputs of the 4th, 7th, 10th and 13th layers) to the 16th, 19th, 22th and 25th layers. Hence the resultant model consists of a shrinking part which aims to capture context, a symmetric expansive part that gradually combines features to enable a precise localization, and a skip connection part that alleviates the vanishing gradient problem and improves the speed of the optimization convergence facilitating the training.

Liu et al. (2020) present a multi-scale feature-based CNN model, called M2DCNN not only to segment WMH, but also to distinguish them from ischemic stroke lesions. M2DCNN contains two symmetric U-shaped subnets that produce multi-scale features through the inclusion of dense and dilated blocks. The former helps reducing the number of training parameters and alleviate the gradient vanishing problem. The latter helps enlarging the receptive fields of the convolution blocks without reducing the feature map size. M2DCNN uses a loss function based on the Dice coefficient.

Hong et al. (2020) present a deep-learning architecture that concatenates two U-Net CNN models that use 3 × 3 kernels in their convolutional layers. The first U-Net consists of four down-sampling and four up-sampling convolutional layers, and generates WMH priors from brain-extracted co-registered T1W and FLAIR images. These WMH candidates, together with the brain-extracted co-registered T1W and FLAIR images, are input to the second U-Net, consisting of two down-sampling and two up-sampling convolutional layers, which reduces the false-positives.

3.5.2. Unsupervised WMH segmentation methods

Damangir et al. (2017) developed an unsupervised method that statistically defined WMH based on the one-tailed Kolmogorov-Smirnov test (Gail and Green, 1976).

Zhan et al. (2015) present an unsupervised WMH segmentation method for T1 and FLAIR data. The T1 image is, first, segmented into different normal tissues, among which regions of white matter and grey matter are combined to provide a region of interest that is subsequently mapped to the FLAIR image. Secondly, the authors calculated the z-score of the intensities in the ROI and defined a threshold to find the abnormalities in normal tissues. They then employ a level set method to improve the preliminary thresholding-based segmentation results and extracted the WMH. The authors pointed out that LGDF energy aided to obtain precise segmentation results compared to other level set methods that used global intensity information.

Bowles et al. (2017) propose a method built upon previous work by the same authors, which can detect abnormally hypertensive regions on FLAIR, disregarding the underlying pathology or location by combining image synthesis, Gaussian mixture models, and one-class support vector machines trained only on healthy tissue.

Valverde et al. (2017) integrate a partial-volume tissue segmentation with WM outlier rejection and filling, combining intensity, probabilistic and morphological prior maps in a pipeline consisting of five steps. These are: 1) Register three statistical a-priori tissue atlases (CSF, GM and WM) and a brain structure atlas to the patient space; 2) Perform atlas-based 5-tissue segmentation on the T1-weighted image; 3) Detect and refill WM outliers as normal-appearing WM based on the registered a-priori and hyper-intense FLAIR maps if available (using the segmentation from step 2); 4) Re-estimate (again) the 5-tissue classes; and 5) Reassign intermediate volume maps into CSF, GM and WM using both neighbour and spatial prior information.

Wang et al. (2015) model the WMH in FLAIR images as having either Gumbel or Fréchet histogram distributions (see Table 1) and compare the results of their algorithm with those from applying a trimmed likelihood estimator. Although results were not accurate for all degrees of lesion loads authors recommend the principle, especially using the Fréchet distribution, due to its simplicity, for studies of ageing and vascular dementia, likely to include subjects with moderate-to-high lesion load.

Atlason et al. (2019) present an autoencoder Segmentation Auto-Encoder (SegAE) consisting on a CNN with architecture similar to that of U-Net with an additional linear layer and parameter constraints to perform linear unmixing. In this model, down- and up-sampling are performed with strided convolutions of 3D kernels of size 2 × 2 × 2 and skip connections are added between activations of the same spatial resolution from the down-sampling to the up-sampling paths. The pure tissue, WMH and cerebrospinal fluid masks obtained during the segmentation were used as priors for the N4 algorithm. Thus, the b1 inhomogeneities are corrected during the training phase and segmentation takes place in presence of inhomogeneity artifacts.

Rachmadi et al. (2020) present an unsupervised segmentation method called Limited One-Time Sampling Irregularity Map (LOTS-IM). This method generates an irregularity map (IM) that represents all voxels as irregularity values ranging from 0 to 1 with respect to the ones considered “normal” based on the original FLAIR texture information. The scheme hierarchically samples a limited number of target squared patches (i.e., 2D patches of 1 × 1, 2 × 2, 3 × 3 and 4 × 4) from a non-overlapping grid of source patches of the same size on each brain-extracted FLAIR image slice, assigning an irregularity value to each source patch. The final irregularity map is generated by blending the hierarchically generated irregularity maps, penalizing the result using the original FLAIR intensities, and normalizing the final values between 0 and 1. The WMH are obtained by thresholding the irregularity map.

Fiford et al. (2020) examined an unsupervised segmentation method Bayesian Model Selection (BaMoS) (Sudre et al., 2015), which models the data as a multivariate mixture of Gaussians, further optimized using the expectation-maximization algorithm. It uses an initial outlier map derived after convergence of the initial Gaussian mixture model to enhance sensitivity, as proposed by Sudre et al. (2017). Newly, this study incorporates a two-threshold selection of the candidate regions and selects the WMH clusters after applying connected component analyses twice, considering first 18 neighbourhoods, and then 6 neighbourhoods, to avoid discarding regions where artefacts and true WMH are present.

3.6. Additional publications evaluating WMH segmentation methods

Dadar et al. (2017b) compare the performance of 10 different linear and non-linear supervised classification methods segmenting WMH in brain scans from 201 subjects from four different datasets. The methods evaluated are: Naïve Bayes, Logistic Regression, Linear and Quadratic Discriminant Analyses, k-NN, decision trees, RF, AdaBoost, SVM and Bagging. Out of these methods, RF was the best performer.

Kujuf et al. (2019) comparatively evaluate 20 methods presented at the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) WMH segmentation challenge in 2019. All algorithms are trained with 60 image datasets acquired in 3 different MR scanners, and evaluated on 110 image datasets from 5 MR scanners. All image data are composed of T1-weighted and FLAIR brain-extracted, bias-corrected and co-registered images from patients with various degrees of age-related neurodegeneration and presenting different vascular pathologies. However, the WMH volume distribution across the dataset is skewed towards low-to-medium WMH burden. From the 20 methods evaluated 14 are neural network approaches, four involve RF, one uses logistic regression, and one a three-level Gaussian mixture model. The evaluation combines the results from five similarity
metrics: DSC, a modified Hausdorff distance (95th percentile), absolute percentage volume difference, sensitivity (recall), and F1-score for individual lesions. The top-ranked methods use ensembles of neural networks.

Rachmadi et al. (2017) compare the performance of two conventional machine learning classifiers (i.e. support vector machine (SVM) and RF) with the performance of three deep learning algorithms, namely the deep Boltzmann machine, convolutional encoder network and a CNN dual-pathway architecture developed specifically for brain lesion detection, for segmenting WMH on brains displaying only mild or no vascular pathology. The results from these five supervised machine-learning methods are also compared with the results from the unsupervised lesion growth algorithm (LGA) of the Lesion Segmentation Tool (LST) publicly available. The evaluation uses FLAIR and T1-weighted images from 20 subjects randomly selected from the ADNI database (http://adni.loni.usc.edu/), acquired in three consecutive years, and for which ground truth WMH segmentations from two different analysts were available. Authors adapted (and/or implemented) configurations that were reported to give the highest WMH segmentation accuracy in previous works. For SVM and RF this study evaluates several combinations of feature vectors with lengths ranging from 44 to 4000, all reported previously having generated results from similar quality. The optimum threshold that defines the boundaries of the probabilistic WMH segmentations differed across methods. Differences in methods’ performance depending on the WMH burden prompted authors to conclude that deep-learning methods, in general, performed better than the two conventional machine learning classifiers (i.e. SVM and RF), being the patch-based CNN configuration the best approach only for scans with low burden of WMH.

3.7. Segmentation descriptive quality

We analyzed the descriptive segmentation method qualities using the scale developed by Byrne et al. (2016). The segmentation descriptive quality (SDQ) is rated on a three point scale: 1 - indicates description of the segmentation method; 2 - indicates explanation of the segmentation method, but no description of how each step is applied; and 3 – indicates full explanation of how the segmentation method proposed is applied. 23/37 studies scored 3.

3.8. Processing time of segmentation methods

Processing time of machine learning based segmentation methods refers to: 1) time taken to load the image; and 2) time taken for application of automatic segmentation algorithm (Cruz et al., 2017). Only 9/37 studies reported the processing time of segmentation method (Ling et al., 2018; Rachmadi et al., 2018, 2020; Manjón et al., 2018; Dadar et al., 2017a; Jiang et al., 2018; Qin et al., 2018; Griffanti et al., 2016; Atlason et al., 2019). Out of these nine studies, four reported the time consumed by the segmentation per image. It ranged from 0.03 s to 9 s per MRI. The method proposed by Qin et al. (2018) consumed considerably less time (i.e., 0.03 s per image) compared to the rest.

3.9. Methods evaluation

The proposed method was cross validated with leave-one-subject-out evaluation in six studies (Stone et al., 2016; Moeskops et al., 2018; Jiang et al., 2018; Knight et al., 2018; Li et al., 2018; Wu et al., 2019a). Fourteen studies described a method for false positive removal (Sadre et al., 2017; Manjón et al., 2018; Stone et al., 2016; Moeskops et al., 2018; Rincon et al., 2017; Bowles et al., 2017; Li et al., 2018; Ghafoorian et al., 2017; Roy et al., 2015; Zhan et al., 2017; Hong et al., 2020; Fiford et al., 2020; Rachmadi et al., 2020; Ding et al., 2020). Fazekas visual rating scale was used in the validation of the results in nine studies (Rachmadi et al., 2018; Qin et al., 2018; Guerrero et al., 2017; Griffanti et al., 2016; Ling et al., 2018; Moeskops et al., 2018; Jiang et al., 2018; Bowles et al., 2017; Rachmadi et al., 2020).

Of the 37 studies, 29 studies evaluated the performance of their WMH segmentation method using the Dice Similarity Coefficient (DSC) among other metrics that measure spatial concordance between the results of the method proposed and reference segmentations. Average DSC values ranged from 0.538 to 0.91 (Table 2). The unsupervised segmentation method proposed by (Damangir et al., 2017) reported the highest average DSC value for WMH segmentation (DSC ranging from 0.85 up to 0.91), followed by the also unsupervised scheme proposed by Wang et al. (2015) (DSC ranging from 0.81 to 0.84), and the k-NN scheme proposed by Jiang et al. (2018) (UBO detector, DSC 0.85). The Bland Altman plot (Martin Bland and Altman, 1986) was used in five studies to analyse the volumetric agreement between the method’s result and manual segmentation (Qin et al., 2018; Guerrero et al., 2017; Ling et al., 2018; Sadre et al., 2015; Fiford et al., 2020). Only four studies validated their method in longitudinal samples (Sadre et al., 2017; Jiang et al., 2018; Rachmadi et al., 2018, 2020), and eight performed an additional validation (i.e., to the traditional comparison against reference standard measurements) using clinical parameters (Guerrero et al., 2017; Jiang et al., 2018; Qin et al., 2018; Rachmadi et al., 2018, 2020; Schirmer et al., 2019; Wu et al., 2019a; Fiford et al., 2020). Comparison with other methods’ performance was done in 29/37 studies. The reference algorithms for excellence were the Lesion Growth Algorithm (LGA) and the Lesion Prediction Algorithm (LPA), both unsupervised methods from the Lesion Segmentation Tool (LST) for SPM (https://www.applied-statistics.de/lst.html).

4. Discussion

In the five-year period evaluated, 37 studies proposed new, or adapted and re-purposed existing approaches, for segmenting WMH of presumed vascular origin from brain MRI. Of these, only 10 were unsupervised. Within the last two years, considerable efforts have been put into developing deep learning WMH segmentation methods particularly based on CNN architectures that have demonstrated success in similar tasks. From the supervised algorithms, 37 % used state-of-the-art CNN and the rest used either conventional machine-learning algorithms, the k-NN algorithm or logistic regression models. Despite the high accuracy usually reported by CNN algorithms, those reviewed do not outperform, in terms of spatial agreement with reference segmentations, the more traditional clustering (i.e. k-NN) and logistic regression supervised methods or the unsupervised methods published in this period. Probably the simplicity and strong priors of the k-NN and logistic regression methods make them easier to train with less data, and are less susceptible to overfitting when training data is limited, compared to the deep-learning schemes. The fact that most of these methods give probabilistic outputs, may be helpful in quantifying marginally pathological tissues like dirty-appearing white matter, and help in the characterisation of ill-defined WMH boundaries. However, it also conspire against their evaluation since these probabilistic results need to be binarised for comparison with manually-derived segmentation binary masks. Quality of reporting has a considerable effect on studies’ value. Poor reporting of the pre-processing and segmentation methods’ steps and lack of availability of the code significantly affects the applicability of various studies included in this review.

We evaluated the validity and accuracy of the segmentation methods reviewed. We refer to validity as the extent to which these algorithms measure what they intended to beyond the data used to develop (i.e., train) and validate them, thus including the applicability to other data. Most studies ignore the issues pertaining to validity and focus only on accuracy of their algorithms. The validity of the proposed segmentation methods was not always clear, mainly due to the different sources of bias in the reference used to evaluate the algorithms (i.e. observer bias in manually-delineated ground truth), the sample selection, and the data source (i.e. mainly from a single protocol and / or acquired from scanners with the same field strength). Many studies exhibited observer bias,
either in training or in evaluating their algorithms, as manual outlines of WMH are always affected by the observer’s perception in recognising a true lesion from an artefact and are influenced by the observer’s experience and ability in delineating the lesion boundaries on MR images. Moreover, reference segmentations are generally obtained by manually refining a semi-automatic segmentation result, obtained generally by thresholding followed by a region-growing algorithm. Selecting the optimum threshold to segment WMH from FLAIR MRI can also be a source of bias (Valdés Hernández et al., 2010). Additionally, not all the studies included were absent of having data selection bias, which can facilitate overfitting if this is not properly addressed. Data augmentation helps reducing overfitting and increasing the number of the training data. However, effects of bias cannot be balanced-out by increasing the sample size or by repetition (Schmidt and Factor, 2013).

It is important to describe the target population, which informs the individuals for whom the results of the study are intended to apply. It can be inferred from the data used in the method development. Studies that validated their methods on a dataset different from the one used to develop it, in terms of clinical and image acquisition characteristics, obtained a lower spatial agreement in this validation dataset (Roy et al., 2015; Wang et al., 2015; Atlason et al., 2019) (Table 2). Many of the studies included analysed the WMH load in the sample, only expressing that it “was representative of the whole load of WMH burden”. However, representativeness does not mean “balanced”: unbalanced data biases the results in favour of the dominant data subgroup – generally patients with medium-to-large WMH burden. Also, many studies did not explain the rationale followed for data selection. For instance, patients with mild cognitive impairment, Alzheimer’s disease, and normal cognition were included in the same study without explaining the selection criteria and relevance for the main objective of the study, i.e. segmenting WMH. For sample sizes like the ones observed in the majority of studies included (e.g. n=100), cognitive status is not a proxy for WMH load (Damangir et al., 2017). It is, therefore, difficult to decide for which level of severity of a particular condition or for which neurological condition the segmentation method had performed well and, therefore, would be recommended.

Many of the included studies used the open access datasets or the datasets provided for the different Lesion Segmentation Challenges (Reinke et al., 2018), Mendelson et al. (2017) pointed out that using an open access dataset to evaluate the performance of a segmentation method introduces selection bias (Mendelson et al., 2017). It indeed is practical, cost effective and allows comparability between methods, but only within the context of the dataset used, especially in the case of supervised methods. Hence, segmentation studies can suffer from limited high-quality data, which is required for training, and poorly labelled region of interests (Challen et al., 2019). The full value of a large dataset depends on the accuracy and completeness of the data collection, which is expensive and time consuming. The use of a limited dataset in cross-validation can falsely show high performance. To evaluate the performance of a segmentation method, large collections of image data are required. Data augmentation and high quality synthetic data can help addressing this need.

Segmenting a medical image is a laborious task. In general, it requires two main steps: 1) image pre-processing; and 2) segmentation (Jude Hemanth and Anitha, 2012). Pre-processing steps generally involve registration, brain extraction, intensity inhomogeneity correction, noise reduction and intensity normalisation (García-Lorenzo et al., 2015). Task-unrelated pathologies (e.g., stroke lesions in SVD neuroradiological features) or imaging artefacts would affect the segmentation algorithm, their identification should be part of the segmentation framework. Main objectives of pre-processing are removal of noise and confounding features, and improving image quality. Not reporting all these steps can be interpreted as they not being necessary or part of the segmentation framework, affecting its reproducibility. Good reporting quality is extremely important, to ensure that accurate and trustworthy information is obtained from the published studies (Samuel et al., 2016). Quality of reporting research studies needs to be improved by following the guidelines outlined by various organisations (Reporting guidelines | The EQUATOR Network, 2020). Institutional strategies to stimulate high quality peer-review to ensure peer-reviewed published reports are in compliance with ICMJE guidelines would be also helpful.

Accuracy of the segmentation methods evaluated in this review refers to their ability to distinguish WMH from normal appearing white matter or other pathological features of similar appearance, as well as a reference or “ground truth” segmentation manually generated by experts. Accuracy was estimated with Bland Altman plots, Jaccard Index, intra class correlation coefficient, true and false positives and negatives and DSC. All these measures have advantages and drawbacks when applied to this context, as none of them alone gives the necessary information about the precision and further applicability of the results in a clinical context. For example, the Bland-Altman plot per-se only allows volumetric comparison between the target and reference methods. The Jaccard Index and DSC are equivalent and they reflect the spatial agreement between the two masks, but do not express how well the algorithm identified the true WMH and/or excluded the non-WMH voxels, as the true positives and negatives are given by other measurements (e.g., true positive fraction, true negative fraction, positive predicted values, false negative/positive fractions). Finally, the correlation coefficient between quantitative WMH volumes and clinical visual ratings (e.g. Fazekas scores), although of clinical use, only gives a gross estimate of how close to the neuroradiological assessment the segmentation is. Most of the papers included did not analyse the results of these metrics combined. It reinforces the claim by Pellegrini et al. (2018) that the relevant literature of computational segmentation algorithms is still insufficiently intertwined with the clinical world. We believe this depends, at least in part, on a misalignment of targets and methods. The computer scientists’ community still aims primarily for algorithm novelty and reaching high levels of precision, experimenting with methods largely inspired by recent developments in the field of computer vision. The clinical research community, on the other hand, aims to verify associations (e.g., biomarkers for outcome, effect of drugs vs. placebo) with clinically relevant features that reflect in an improvement in patient outcomes using statistical models. A combination of methods and aims is, therefore, of great importance. The fact that 22% of the studies analysed incorporated a clinical validation to their scheme is encouraging.

Only 26% of the studies included in the review reported the processing time of the proposed segmentation algorithm. Reporting the processing time of the segmentation method could also aid in its further translation to clinical practice, by highlighting its speed or the need of optimising its current implementation. Despite its importance, translating research into clinical practice is challenging. Aside from simply demonstrating superior efficacy, new technologies entering the medical field must also integrate with current practices and be effective in an individual case basis. Kristensen et al. (2015) have reported that it takes more than a decade to implement research results in clinical practice. The research required for this ‘personalised’ medicine would only be possible through summarising and integrating enormous quantities of medical information. This review has shown that this is still unachieved.

This review systematically extracted, synthesised, critically appraised and presents information about a highly active research field. Its main strengths are: 1) careful selection of relevant studies amongst a vast number of initial candidates resultant from the search; 2) identification of the possible useful bias in the studies; and 3) synthesis of the contributions of the included papers. Limitations are that we only included the articles published in English language for which we have full access: we may have missed articles published in languages other than English or other articles for which we could not access the full text. Also, there might be other relevant papers missing as a result of incongruences between the search terms and the article keywords or indexing in the databases. By excluding articles published in conference proceedings it is possible that
promising WMH segmentation methods could have been excluded.

5. Conclusion and future works

Despite the increasing popularity and high accuracy of CNN schemes applied to WMH segmentation, we found no evidence to favour their promising WMH segmentation methods could have been excluded.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.medim.2021.101867.

References

Advanced Normalization Tools (ANTs) - SourceForge.net. (2020). Retrieved 18 June 2020, from https://sourceforge.net/projects/advants/files/.
Aksjudov, T.N., Nabudzi, I., Makeyte, M., Scanlon, C., Hehir, S., Casey, H., Ambati, S., Kenney, J., O’Donoghue, S., McDermott, E., Killiany, R., Dockery, P., McDonald, C., Hallahan, B., Cannon, D., 2018. A comparative study of segmentation techniques for the quantification of brain subcortical volume. Brain Imaging Behav. 12 (6), 1678-1695. https://doi.org/10.1007/s11682-018-9635-y.
Anbeek, P., Vincen, K.L., Van Osch, M.J., Bisschops, R.H., Van der Grond, J., 2004. Online version, at doi: https://doi.org/10.1016/j.compmedimag.2021.
Appendix A. Supplementary data

Debette, S., Markus, H., 2010. The clinical importance of white matter hyperintensities in Alzheimer disease. Neuroimage 51 (4), 1250-1259. https://doi.org/10.1016/j.neuroimage.2010.04.017.
De Boer, R., van der Lijn, F., Vrooman, H., Vernooij, M., Ikram, M., Breteler, M., De Deij, R., Koek, M., Polman, C., Stroink, H., 2018. Detection of white matter lesion regions in MRI using SLIC0 and convolutional neural networks. Comput. Med. Imaging Graph. 76, 9-16. https://doi.org/10.1016/j.compmedimag.2019.102085.
Dadar, M., Marzani, J., Missiaglia, K., Anor, C.J., Fonov, V.S., Tartaglia, M.C., Carmichael, O.T., De Carli, C., Collins, D.L., 2017a. Validation of a registration technique for segmentation of white matter hyperintensities in Alzheimer’s disease. IEEE Trans. Med. Imaging 36 (8), 1758-1768. https://doi.org/10.1109/TMI.2017.2693978.
Dadar, M., Marzani, J., Missiaglia, K., Anor, C.J., Vos, F.S., Tartaglia, M.C., Carmichael, O.T., De Carli, C., Collins, D.L., Alzheimer’s Disease Neuroimaging Initiative, 2017. Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. NeuroImage 157, 233-249. https://doi.org/10.1016/j.neuroimage.2017.06.005.
Dalca V., Adrian, et al., 2014. Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. Med Image Comput Assist Interv 17, 773-780. https://doi.org/10.1007/978-3-319-10470-6_96. In press.
De Boer, R., van der Lijn, F., Vrooman, H., Vernooij, M., Ikram, M., Breiter, M., Niessen, W., 2007. Automatic segmentation of brain tissue and White matter lesions in MRI. In: Proceedings of the 2007 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Washington, DC, USA, April 12-16, 2007. doi:10.1109/ISBI.2007.359361.
Debette, S., Markus, H., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 341 (jul26 1). doi:10.1136/bmj.c3666. c3666. c3666.
Descoteaux, M., Ayache, N., 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425-441. https://doi.org/10.1109/TMI.2007.906087.
Dixit, R., O’Connor, J., 2018. Probabilistic segmentation of white matter lesions in MS imaging. NeuroImage 175, 818-835. https://doi.org/10.1016/j.neuroimage.2018.04.011.
Dixon, T., Cohen, A.D., O’Connor, E.E., Karim, H.T., Craimichean, A., Muschelli, J., Lopez, O., Klink, W.E., Azenstein, J.H., Krafty, R., Craimichean, C.M., Tudosorac, D. L., 2020. An improved algorithm of white matter hyperintensity detection in elderly adults. Neuroimage Clin. 25, 102151. https://doi.org/10.1016/j.nic.2019.102151.
Diniz, P.H.B., Valente, T.L.A., Diniz, J.O.B., Silva, A.C., Gattass, M., Ventura, N., et al., 2017. Fast evaluation of segmentation subject to bounds. Siam J. Optim. 26 (3), 1622-1637. https://doi.org/10.1137/16M1080370.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
De Vos, F.S., Stavrakis, A., Barr, C.L., 2008. Artificial intelligence, bias and clinical validity. BMJ Qual. Saf. 28 (3), 231-237. https://doi.org/10.1136/bmjqs-2018-008370.
Diniz, P.H.B., Valente, T.L.A., Diniz, J.O.B., Silva, A.C., Gattass, M., Ventura, N., et al., 2017. Fast evaluation of segmentation subject to bounds. Siam J. Optim. 26 (3), 1622-1637. https://doi.org/10.1137/16M1080370.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
De Vos, F.S., Stavrakis, A., Barr, C.L., 2008. Artificial intelligence, bias and clinical validity. BMJ Qual. Saf. 28 (3), 231-237. https://doi.org/10.1136/bmjqs-2018-008370.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
Davies, D., Martiski, K., Verbiest, A., Zijdenbos, A.P., 2010. Multi-modality probabilistic atlas construction using semi-supervised learning. Med. Image Anal. 14 (2), 157-168. https://doi.org/10.1016/j.media.2009.11.002.
pipeline for detecting small deep white matter hyperintensities in migraineurs. Neuroimage Clin. 18, 638–647. https://doi.org/10.1016/j.nicl.2018.02.033.
Pellegriini, E., Bellerini, L., Hernandez, M., Chappell, F.M., Gonzalez-Castro, V., Anblagan, D., Danso, S., Muñoz-Manjáさえ, J., Job, D., Pernet, C., Maír, G., MaggiLibra, T.J., Trucco, E., Wardlaw, J.M., 2018. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. J. Neuroimaging 28 (4), 519–529. https://doi.org/10.1177/1052620618790594.
Qin, C., Guerero, R., Bowles, C., Chen, L., Dickie, D., Valdes-Hernandez, M., et al., 2018. A large margin algorithm for automated segmentation of white matter hyperintensity. Pattern Recognit. 77, 156–159. https://doi.org/10.1016/j.patcog.2017.12.016.
Rachmadi, M., Valdes-Hernandez, M., Agan, M., Komura, T., 2017. Deep learning vs. conventional machine learning: pilot study of WMHI segmentation in brain MRI with absence or mild vascular pathology. J. Imaging 3 (4), 66. https://doi.org/10.3390/jimaging3040066.
Rachmadi, M.F., Valdes-Hernandez, M.A., DiPirri, C., Komura, T., Alzheimers Disease Neuroimaging Initiative, 2018. Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. Comput. Med. Imaging Graph. 66, 28–43. https://doi.org/10.1016/j.compmedimag.2018.02.002.
Richard, E., Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S., Staring, M., Alzheimer Schlemm Schmidt, R.L., Factor, R.E., 2013. Understanding sources of bias in diagnostic accuracy Reporting guidelines, 2020. Reporting Guidelines | The EQUATOR Network. Retrieved 5 July 2020 from https://reporting-guidelines-equator-network.org/.
Sweeney, E.M., Shinohara, R.T., Shies, N., Mateen, F.J., Chudgar, A.A., Cuzzocreo, J.L., Calabresi, P.A., Pham, D.L., Reich, D.S., Crainiceanu, C.M., 2019. Automated statistical inference for segmentation with applications to multiple sclerosis lesion segmentation in MRI. Neuroimage Clin. 22, 101772 https://doi.org/10.1016/j.nicl.2019.101772.
Tustison, J., Nicholas, et al., 2010. NITK: Improved N3 Bias Correction. IEEE Transactions on Medical Imaging 29 (6), 1310–1320. https://doi.org/10.1109/TMI.2010.2049690.
Van Opbroek, A., Ikram, M., Vernooij, M., de Bruijne, M., 2015a. Transfer learning improving supervised imaging: a review of transfer learning imaging protocols. Neuroimage Clin. 24, 1018–1020. https://doi.org/10.1016/j.nicl.2015.09.011.
Van Opbroek, A., Ikrum, M., Vernooij, M.W., de Brujin, M.A., 2015b. Transfer learning improving supervised imaging: a review of transfer learning imaging protocols. Neuroimage Clin. 24, 1018–1020. https://doi.org/10.1016/j.nicl.2015.09.011.
Wang, R., Li, C., Wang, J., Wei, X., Li, Z., Zhu, Y., Zhang, S., 2015. Automated segmentation and volumetric quantification of white matter hyperintensities on fluid-attenuated inversion recovery images using the extreme value distribution. Neuroradiology 57 (3), 307–320. https://doi.org/10.1007/s00234-014-1466-4. Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12 (8), 822–838. https://doi.org/10.1016/S1474-4422(13)70124-8.
Wardlaw, J., Valdes-Hernandez, M., Muñoz-Manjáさえ, M., 2015. What are White Matter Hyiperintensities Made Of. J. Am. Heart Assoc. 4 (6). https://doi.org/10.1161/JAHA.115.000930.
Wu, D., Albert, M., Soldan, A., Pettigrew, C., Oishi, K., Tomogane, Y., Ye, C., Ma, T., Wu, Y., Liu, Y., 2013. Adaptively weighted large margin classifiers. J. Comput. Graph. Geom. Med. 16 (6), 805–856. https://doi.org/10.1109/jcmg.2013.623278.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 987–990. https://doi.org/10.1109/42.811270.
Van Opbroek, A., Ikrum, M., Vernooij, M., de Brujin, M.A., 2015a. Transfer learning improving supervised imaging: a review of transfer learning imaging protocols. Neuroimage Clin. 24, 1018–1020. https://doi.org/10.1016/j.nicl.2015.09.011.
Van Opbroek, A., Ikrum, M., Vernooij, M.W., de Brujin, M.A., 2015b. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Med. Image Anal. 24 (1), 245–254. https://doi.org/10.1016/j.media.2015.06.016.
Viole, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24 (2), 137–154. https://doi.org/10.1023/A:1007920415769.
Wu, D., Albert, M., Soldan, A., Pettigrew, C., Oishi, K., Tomogane, Y., Ye, C., Ma, T., Wu, Y., Liu, Y., 2013. Adaptively weighted large margin classifiers. J. Comput. Graph. Geom. Med. 16 (6), 805–856. https://doi.org/10.1109/jcmg.2013.623278.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 987–990. https://doi.org/10.1109/42.811270.
Van Opbroek, A., Ikrum, M., Vernooij, M., de Brujin, M.A., 2015a. Transfer learning improving supervised imaging: a review of transfer learning imaging protocols. Neuroimage Clin. 24, 1018–1020. https://doi.org/10.1016/j.nicl.2015.09.011.
Van Opbroek, A., Ikrum, M., Vernooij, M.W., de Brujin, M.A., 2015b. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Med. Image Anal. 24 (1), 245–254. https://doi.org/10.1016/j.media.2015.06.016.
Viole, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24 (2), 137–154. https://doi.org/10.1023/A:1007920415769.
Wu, D., Albert, M., Soldan, A., Pettigrew, C., Oishi, K., Tomogane, Y., Ye, C., Ma, T., Miller, M.I., Mori, S., 2019b. Multi-atlas-based detection and localization (MADL) for location-dependent quantification of white matter hyperintensities. Neuroimage Clin. 22, 101772 https://doi.org/10.1016/j.nicl.2019.101772.
Zhan, T., Liu, Z., Xiao, L., Zhan, Y., Wei, Z., 2020. Automatic method for white matter lesion segmentation based on T1-fluid-attenuated inversion recovery images. J. Comput. Med. Imag. Vis. 9 (4), 447–455. https://doi.org/10.1016/j.jcmiv.2020.03.011.
Zhang, T., Yu, Z., Zheng, Y., Zhan, Y., Xiao, L., Wei, Z., 2017. Multispectral spatial segmentation framework for white matter lesions in multi-sequence magnetic resonance images. Biomed. Signal Process. Control 31, 52–61. https://doi.org/10.1016/j.bspc.2016.09.003.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57. https://doi.org/10.1109/42.908.284.
Zg, L., Delbarre, K., Chosker, P., et al., 2011. Impact of white matter lesions on physical functioning and fall risk in older people. Stroke 42 (7), 1806–2090. https://doi.org/10.1161/strokeaha.110.601360.