Neuroradiology
https://doi.org/10.1007/s00234-023-03256-0

REVIEW

A systematic review of (semi-)automatic quality control of T1-weighted MRI scans

Janine Hendriks[1] · Henk-Jan Mutsaerts[1] · Richard Joules[2] · Óscar Peña-Nogales[3] · Paulo R. Rodrigues[3] · Robin Wolz[2,4] · George L. Burchell[5] · Frederik Barkhof[1,6] · Anouk Schrantee[7]

Received: 28 September 2023 / Accepted: 16 November 2023

Abstract

Purpose Artifacts in magnetic resonance imaging (MRI) scans degrade image quality and thus negatively affect the outcome measures of clinical and research scanning. Considering the time-consuming and subjective nature of visual quality control (QC), multiple (semi-)automatic QC algorithms have been developed. This systematic review presents an overview of the available (semi-)automatic QC algorithms and software packages designed for raw, structural T1-weighted (T1w) MRI datasets. The objective of this review was to identify the differences among these algorithms in terms of their features of interest, performance, and benchmarks.

Methods We queried PubMed, EMBASE (Ovid), and Web of Science databases on the fifth of January 2023, and cross-checked reference lists of retrieved papers. Bias assessment was performed using PROBAST (Prediction model Risk Of Bias ASsessment Tool).

Results A total of 18 distinct algorithms were identified, demonstrating significant variations in methods, features, datasets, and benchmarks. The algorithms were categorized into rule-based, classical machine learning-based, and deep learning-based approaches. Numerous unique features were defined, which can be roughly divided into features capturing entropy, contrast, and normative measures.

Conclusion Due to dataset-specific optimization, it is challenging to draw broad conclusions about comparative performance. Additionally, large variations exist in the used datasets and benchmarks, further hindering direct algorithm comparison. The findings emphasize the need for standardization and comparative studies for advancing QC in MR imaging. Efforts should focus on identifying a dataset-independent measure as well as algorithm-independent methods for assessing the relative performance of different approaches.

Keywords Systematic review · Quality control · Structural MRI · Rule-based learning · Machine learning · Deep learning

✉ Janine Hendriks
j.m.hendriks@amsterdamumc.nl

1 Department of Radiology and Nuclear Medicine, Amsterdam UMC, Location VUmc, PK -1, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands

2 IXICO Plc, London EC1A 9PN, UK

3 QMENTA, Barcelona 08009, Spain

4 Imperial College London, London SW7 2BX, UK

5 Medical Library, Vrije Universiteit Amsterdam, Amsterdam 1081 HV, The Netherlands

6 Queen Square Institute of Neurology and Centre for Medical Image Computing, University College London, London WC1N 3BG, UK

7 Department of Radiology and Nuclear Medicine, Amsterdam UMC, Location AMC, Amsterdam 1105 AZ, The Netherlands

Introduction

Significant advances have been made in the realm of medical image analysis in the past few decades [1]. Imaging biomarkers derived from advanced imaging techniques such as magnetic resonance imaging (MRI) data are used to characterize normal development [2], disease [3], and the effects of disease-modifying therapies [4]. T1-weighted (T1w) MRI scans, for example, depict the anatomical arrangement of gray matter, white matter, and cerebrospinal fluid, providing valuable insights into the brain's structural composition or pathology. However, before any image analysis workflow can be used, the quality of the MRI scans has to be ensured. Scan quality can be degraded by artifacts, which are unexpected or artificial image irregularities that are not related to anatomical or physiological

Published online: 04 December 2023

abnormalities, but that can arise during the imaging process, such as blurring, ghosting, and aliasing. These artifacts can lead to low statistical power or erroneous conclusions.

Currently, the quality of MRI data is often ensured through manual quality control (QC), which traditionally entails visual inspection of every individual scan of a dataset by an expert rater, from which those showing insufficient data quality are excluded. This manual QC is time-consuming and prone to variability. Undesired variability may arise from inter- and intra-rater differences, such as training, experience and fatigue [5]. Additional concerns are that subtle artifacts stemming from improper choice of acquisition parameters may be too subtle to be detected by the human eye [6], or that differences in scanner vendor can introduce variability [7]. These drawbacks of manual QC create great difficulty in defining objective exclusion criteria.

Furthermore, the acquisition of very large datasets across multiple scanning sites [8–10] needed for clinical trials introduces additional concerns. Such large datasets make individual inspection of each image resource-intensive and add the possibility of intra-site/rater variability. Therefore, there has been a great interest in the development of automated QC tools. Over time, several of these (semi-)automated QC algorithms have been created. Fully automated QC algorithms classify MRI scans without human involvement, while semi-automated QC algorithms require some level of human decision-making during the process, such as manually changing a threshold. These (semi-)automated QC algorithms not only make use of thresholds but can also employ classifiers based on classical machine learning or deep learning approaches to categorize MRI scans. Some of the algorithms focus on raw, unprocessed MRI scans [11, 12] whereas others focus on their derivatives in the form of processed scans (e.g., segmentations) or statistics (e.g., regional volumetrics) [13, 14]. However, despite their common goal of QC, these algorithms differ in their outcome parameters, and their application.

In light of the multitude of available tools, determining the most suitable choice can be challenging. Therefore, this review provides an overview of automated QC algorithms and software packages specifically designed for scrutinizing raw structural T1w MRI scans. We focus on whole brain, standard resolution T1-weighted MRI scans, typical of those ubiquitously employed in clinical trials and clinical practice. Our main objective is to identify the distinctions among these QC algorithms in terms of features of interest, performance, outcome metrics, and type of data used as a benchmark.

Methods

Literature screening for this review was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines [15–17] and was registered with the Prospective Register of Systematic Reviews (PROSPERO) database under number CRD42023391301.

Data sources and searches

A systematic search was conducted to identify algorithms for (semi-)automatic quality control of structural MRI scans. Formal methods for literature search, selection, quality assessment, and synthesis were used, according to the PRISMA Guidelines [15–17]. The systematic search was performed in the following databases: PubMed, EMBASE.com, and Clarivate Analytics/Web of Science Core Collection. The timeframe within the databases was from inception to 5th January 2023, and the initial search was conducted by GLB and JH. The search included keywords and free text terms for (synonyms of) "artificial intelligence" or "machine learning" combined with (synonyms of) "quality control" combined with (synonyms of) "Magnetic Resonance Imaging" combined with (synonyms of) "neuroscience." A full overview of the search terms per database can be found in the supplementary information (Supplementary Tables 1—3). No limitations on date or language were applied in the search. Additionally, reference lists of included articles were manually screened to identify additional articles.

Article inclusion and exclusion criteria

Articles were included in this review if the following inclusion criteria were met: (1) involved the performance of (semi-)automatic quality control, (2) evaluated raw T1-weighted human brain MRI scans, and (3) published in English as original research in peer-reviewed journals or conference proceedings (conference abstracts and posters excluded). Articles involving QC of the output of preprocessing pipelines, as well as articles simply applying existing QC algorithms, were excluded.

Data synthesis and analysis

To assess the eligibility of the selected articles, two independent reviewers (JH and AS) reviewed all abstracts from the database searches and retrieved full-text articles for further review. Any discrepancies were resolved through consensus. Finally, one reviewer (JH) read the retrieved articles for final article selection and quality assessment. The bibliographies of the retrieved full-text articles were manually searched for additional publications. For quality assessment, the PROBAST (Prediction model Risk Of Bias ASsessment Tool) [18] method was chosen, being commonly used for assessing the risk of bias and applicability of prediction

model studies. In this tool, the risk of bias is defined to occur when shortcomings in the study design, conduct, or analysis lead to systematically distorted estimates of the model's performance. Concerns regarding applicability of an article to the review question can arise when the population, predictors, or outcomes differ from those specified in the review question.

Data extraction

All full-text articles that met the inclusion criteria were assigned into one of the following three categories: "Rule-based," "Classical Machine learning," and "Deep learning", based on the classification method used in the described algorithm. Rule-based algorithms establish thresholds for predefined quality features, while classical machine learning algorithms utilize a classifier to differentiate two groups based on an empirically established threshold from predefined quality features. Deep learning algorithms differ from those two groups as they do not rely on predefined quality features and use a classifier to determine the groups. All articles were assessed by one rater, and technical information and features of the algorithms including used datasets, age range of included participants, artifact presence, benchmark, QC result, and performance measures were extracted. When possible, missing performance measures were calculated manually from data available in the articles.

Results

Search results

The electronic search yielded 268 hits from PubMed, 496 from EMBASE, and 252 from Web of Science, amounting to a total of 1016 hits (Fig. 1). After removing duplicates, 605 unique articles were identified. After title and abstract review, 576 articles were excluded, and 29 were sought for retrieval of which one could not be retrieved. Of the 28 articles that underwent full-text review, 12 were excluded from further quality assessment because of (1) assessing the quality of already processed (rather than raw) scans ($N=5$), (2) relying on visual QC only ($N=3$), (3) evaluating previously published algorithms ($N=2$), (4) assessing quality of the file structure ($N=1$), and (5) assessing quality of other MR sequences ($N=1$). Cross-reference searching of the included articles resulted in the identification of eight more articles, of which six underwent full-text review of which four were

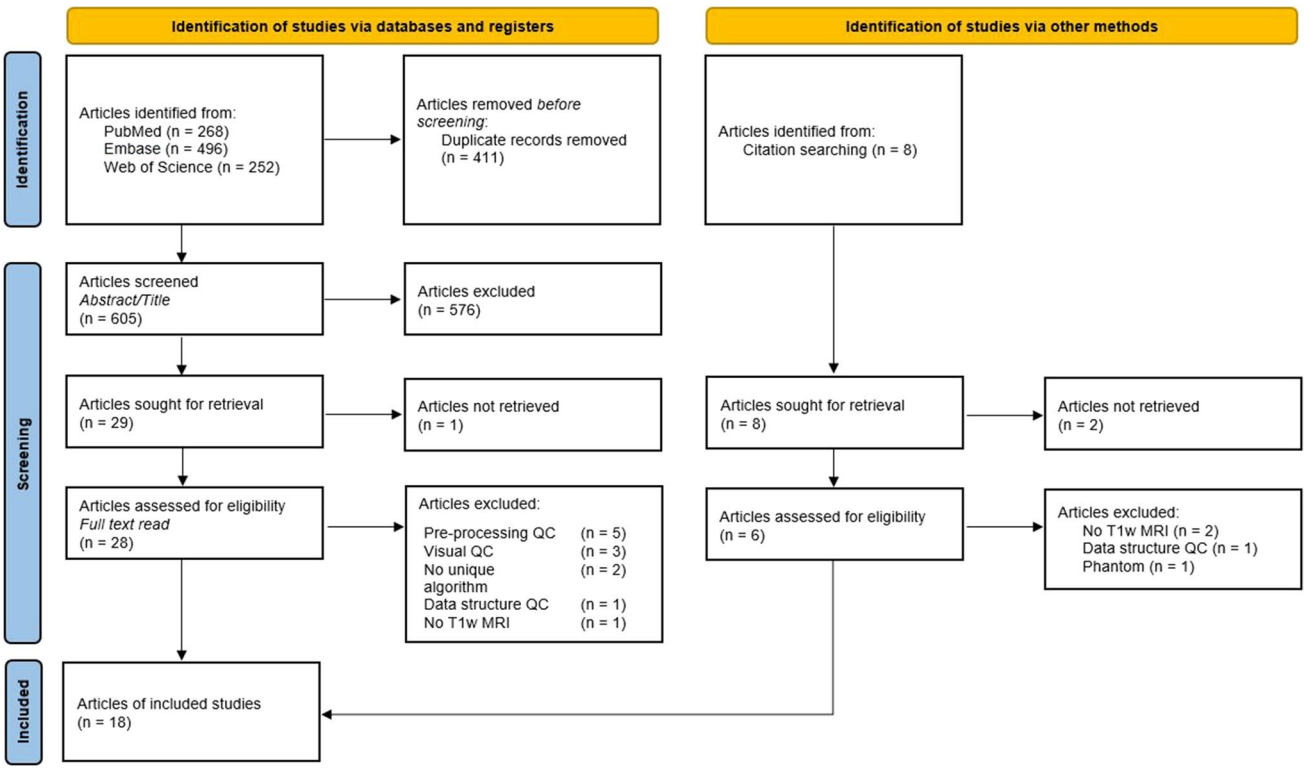

Fig. 1 PRISMA 2020 flow diagram (adapted from: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71

excluded from further quality assessment due to (1) assessing quality of other MR sequences ($N=2$), (2) assessing quality of a file structure ($N=1$), and (3) assessing quality of phantom images ($N=1$). Ultimately, a total of 18 articles were included [8, 11, 12, 19–33]. Relevant information regarding dataset, benchmark, and performance measures is summarized in Table 1. A variety of T1w sequences have been used in the included studies, mostly 3D acquisitions with an inversion recovery spoiled gradient-like protocols (see Supplementary Table 4).

Risk of bias assessment

All included articles showed a low concern regarding applicability of the algorithms (Table 2). For 16 articles, the outcome measures of the articles showed a high risk of bias, since the benchmarks are mostly based on a visual QC (Table 1), which can lead to distorted or flawed assessment in the benchmarks. During visual QC, different raters may interpret a protocol in varying ways. Additionally, they may have reacted uniquely to specific scans, causing the resulting score to, to some extent, reflect the raters' characteristics in addition to solely assessing the quality of the target scan. Examples of factors introducing risk of bias for a visual QC approach are an unclear protocol [11, 22, 26], undisclosed number of raters [11, 20, 26], or QC based solely on the assessment of one rater [8, 24, 30]. Bias in algorithms can rise from unrepresentative or incomplete training data or the reliance on flawed information. Consequently, the analysis of multiple algorithms is susceptible to bias, particularly when a low number of participants [8, 21, 22, 25] was used, potentially resulting in unrepresentative or incomplete training data. Furthermore, the absence of validation [20–22, 26–29, 33] further heightens the risk of bias, as it lacks a safeguard to detect and rectify potential reliance on flawed information within the training data.

Rule-based QC

A total of ten articles were found that utilized a rule-based approach of one or more quality features to evaluate the quality of structural MRI scans [20–22, 24, 26–29, 32, 33]. Eight articles categorize MRI scans into pass and fail groups [21, 22, 24, 26–28, 32, 33], one assessed whether blurring is present [29], and one assessed SNR and ghosting without an overarching score [20]. Five articles [20, 24, 26, 32, 33] evaluated quality features based on the background of the image, arguing that most of the artifactual signal intensities propagate over the image and into the background, which typically corresponds to 40% of the total volume of a structural MRI scan. The other articles [21, 22, 27–29] use foreground-based quality features, arguing that relying solely on the background may not provide a reliable measure of the overall image quality.

Rule-based QC using background

One article used quality features stemming from previous studies trying to assess distortion in scans caused by image compression [33]. In order to assess which of these quality features are most applicable for QC of MRI scans, they applied a large set of quality features, classified into seven feature families, to artificially distorted MRI scans ($N=143$). They found that quality features based on Natural Scene Statistics were the most effective in distinguishing between artificially distorted and undistorted MRI scans.

Subsequent studies investigated features specific for MR artifacts extracted from the background, such as noise and ghosting artifacts in scans ($N=250$) [20]. When these features were compared with white matter SNR and visually assessed ringing artifacts (undisclosed number of raters), very high sensitivity and specificity are found. They determined a QC threshold based on the highest agreement between the features and manual assessment. For the feature assessing ghosting, a validation test was performed, which resulted in similar specificity and sensitivity to the training set.

Subsequent work [26] put forth the argument that SNR measures may not necessarily be sensitive to subject-related artifacts. Instead, they suggest that these artifacts lead to a corrupted noise distribution that can be evaluated using two specific quality features. The first feature assesses the effects of clustered artifacts in the background, and the second feature evaluates both clustered and subtle effects of artifacts in the background ($N=749$; undisclosed number of raters). White et al. [32] continued on this work, by calculating the integral of the voxel intensities as vectors radiating away from the head ($N=6662$; 1 or 2 raters). This feature was compared with two other new features, which capture the frequency characteristics of the noise rippling away from the edge of the head and utilize properties of the line spread function along the edge of the head. In both these studies, visual quality assessment led to a binary pass or fail score, which was either used to determine the cut of values for the quality features [26] or to assess the performance by determining the area under the curve (AUC) [32]. The feature utilizing the line spread function along the edge of the head was reported to perform best [32].

Finally, the LONI QC System [24] is a publicly available QC algorithm for structural T1w scans based on seven features, namely, SNR, signal variance-to-noise variance ratio (SVNR), contrast-to-noise ratio (CNR), contrast of variance-to-noise ratio (CVNR), brain tissue contrast-to-tissue intensity variation (TCTV), full-width-at-half-maximum (FWHM), and center of mass (CoM). In this algorithm, the image

Table 1 Data extraction

Article	Method (# of features)	Dataset				Benchmark	QC result	Performance measures					
		Name	*N*	Age range (y)	Artifact presence (%)			Sensitivity/specificity	Accuracy	PPV/NPV	BA	AUC	Pearson's *R*
[20]	RB (2)	NA	40	n.k	NA	Visual	SNR pass/fail	NA	NA	NA	NA	NA	0.98
		NA	250	n.k	52.4	Visual	Ghosting yes/no	93.2/92.3	NA	NA	92.8*	NA	NA
		NA	50	n.k	NA	Visual	Ghosting yes/no	95.5/96.4	96.0*	95.5*/96.4*	96.0*	NA	NA
[21]	RB (20)	NA	100	25–38	NA	Visual	Ranking	NA	NA	NA	NA	NA	0.92
[22]	RB (20)	ADNI + ABIDE	50	55–90 + 5–64	NA	Visual	Ranking	NA	NA	NA	NA	NA	0.93–0.95
[24]	RB (7)	TRACK-TBI	1569	18–39	1.59	Visual	Pass/fail	85.0/87.0	89	NA	86.0*	0.93	NA
[26]	RB (2)	ADNI + ABIDE	749	55–90	7.21	Visual	Pass/fail	87.2/85.2	NA	NA	86.2*	0.93	NA
[28]	RB (3)	ADNI + NeuroRX	n.k	55–90 + n.k	NA	Simulated	Ranking	NA	NA	NA	NA	NA	NA
[29]	RB (2)	ADNI + NeuroRX	n.k	55–90 + n.k	NA	Simulated	Ranking	NA	NA	NA	NA	NA	0.73
[27]	RB (5)	NeuroRX	n.k	n.k	NA	Simulated	Ranking	NA	NA	NA	NA	NA	NA
[32]	RB (3)	Generation R Wave I	1070	6–111	4.77	Visual	Pass/fail	NA	NA	NA	NA	0.95	NA
		Generation R Wave II	4339	8–122	8.78	Visual	Pass/fail	NA	NA	NA	NA	0.95	NA
		NHGRI	442	5–788	5.2	Visual	Pass/fail	NA	NA	NA	NA	0.88	NA
		GUSTO	811	4–55	54.62	Visual	Pass/fail	NA	NA	NA	NA	0.82	NA
[33]	RB (259)	MNI	143	n.k	NA	Simulated	Pass/fail	NA	NA	NA	NA	NA	NA
[11]	ML (190)	UK Biobank	5816	40–60	1.77	Visual	Pass/fail	91.3/84.0	84.2*	9.3/99.8	87.6	NA	NA
[12]	ML (64)	ABIDE	1102	5–644	30.58	Visual	Pass/fail	NA	76.2	NA	NA	0.73	NA
		DS030	265	21–55	28.3	Visual	Pass/fail	76.9*/67.7*	75.8	94.7*/28.0*	72.3	0.71	NA
[30]	ML (6)	CBDB Sibling Study	1457	18–60	13.04	Visual	Pass/fail	70.1/88.2	79	NA	79.2*	NA	NA
[19]	DL	NA	3770	18–96	43.47	Visual	Pass/fail	79.9/87.1	84.2*	81.9/85.8	83.5	83.5	NA
			2182		40.01	Visual	Medium/good	77.4/65.9	73.9*	83.2/57.8	71.7	71.7	NA
[8]	DL	NA	68	21–61	50	Visual	Pass/fail	NA	90.4	NA	NA	NA	NA
		NA	93	8–666	40.86	Visual	Pass/fail	97.4/85.5*	90.3	82.2/97.9*	91.4*	NA	NA
		NA	22	19–65	9.09	Visual	Pass/fail	100/60.0*	63.6	20.0/100*	80.0*	NA	NA
		NA	20	22–85	30	Visual	Pass/fail	100/71.4*	75	33.3/100*	85.7*	NA	NA
[23]	DL	HBN	200	5–18	50	Visual	Pass/fail	NA	NA	NA	NA	0.99	NA
[25]	DL	NA	32	n.k	50	Visual	Motion yes/no	NA	92	NA	NA	NA	NA
[31]	DL	ABIDE	1064	5–644	12.41	Visual	Pass/fail	77.0/85.0	84	42.0/96.0	91.0*	0.9	NA

n.k., not known; *NA*, not available; *RB*, rule-based; *ML*, machine learning-based; *DL*, deep learning based; *PPV*, positive predictive value; *NPV*, negative predictive value; *BA*, balanced accuracy; *AUC*, area under the curve

*Manually calculated performance measure

Table 2 Risk of bias assessment

Article	ROB			Applicability		Overall	
	Predictors	Outcome	Analysis	Predictors	Outcome	ROB	Applicability
[20]	+	?	-	+	+	-	+
[21]	+	+	-	+	+	-	+
[22]	+	-	-	+	+	-	+
[24]	+	-	+	+	+	-	+
[26]	+	-	-	+	+	-	+
[28]	+	NA	-	+	NA	-	+
[29]	+	-	-	+	+	-	+
[27]	+	NA	-	+	NA	-	+
[32]	-	-	+	+	+	-	+
[33]	+	NA	-	+	NA	-	+
[11]	+	-	-	+	+	-	+
[12]	+	-	+	+	+	-	+
[30]	+	+	+	+	+	+	+
[19]	NA	+	+	NA	+	+	+
[8]	NA	-	-	NA	+	-	+
[23]	NA	-	+	NA	+	-	+
[25]	NA	-	-	NA	+	-	+
[31]	NA	-	+	NA	+	-	+

PROBAST, Prediction model Risk Of Bias ASsessment Tool; *ROB*, risk of bias; *NA*, not applicable. Domain 1: Participants were not relevant and hence disregarded. Domain 2 was disregarded for the algorithms utilizing a deep learning approach, since the predictors are not known a priori in these cases. + (green) indicates low ROB/low concern regarding applicability;—(red) indicates high risk of bias/high concern regarding applicability; ? (yellow) indicates unclear risk of bias/unclear concern regarding applicability

background is used to represent noise. To assess the manual binary classification (one rater) accuracy of each QC feature, the values per feature were changed to a *z*-score, and a threshold was determined based on agreement with visual QC.

Rule-based QC using foreground

Jang et al. [22] argue that not all MRI scans allow for capturing background noise, e.g., when the background area is insufficient to allow a robust analysis. They introduced the Quality Evaluation using MultiDirectional filters for MRI (QEMDIM) algorithm, which uses multidirectional filters to capture quality features. Each image is divided into 16 patches with 20 quality features, which were averaged over the patches. Image quality is determined by calculating the absolute difference between the averaged quality features of the test image and those of a benchmark of undistorted images, using the agreement with visual scores as the threshold. Others [21] later modified QEMDIM such that it does not only provide the absolute quality difference but also if the assessed scan has a higher or lower quality than the benchmark. Additionally, calculation efficiency was improved by omitting patch division and feature averaging. Then, they revalidated the modified QEMDIM score with a visual quality score.

Osadebey et al. [27–29] developed three algorithms to assess MRI scan quality, using foreground features. One algorithm [27] calculates a total quality score as a weighted sum of noise, lightness, contrast, sharpness, and texture details. In a second algorithm [28], three geospatial local entropy features are being extracted from all the slices of MRI scans. In both these algorithms, it was shown that undistorted images have higher quality scores than artificially degraded images. In a more recent study [29], the authors used an average of a sharpness and a contrast quality feature; which also performed good in comparison to a visual rating scale.

Classical machine learning

Three studies applied classical machine learning approaches [11, 12, 30], which all classify MRI scans into either pass

or fail by training them against visual QC results. Quality features based on the background and the foreground of the image have been used.

Pizarro et al. [30] investigated multi-dimensional, non-linear classification to overcome limitations of univariate approaches, including the need for multiple quality features to characterize artifacts from different sources, since a single quality feature has a limited ability to capture details of artifacts in small local regions and cannot capture sufficient information on artifact type and location. Six different features were extracted: three volumetric features (related to contrast, intensity, and tissue class) and three artifact-specific features (related to eye movement, ringing, and aliasing). The MRI scans ($N=1457$) are also visually assessed and classified in either pass or fail by five to nine raters. The features and the visual assessment were fed to a supervised classification algorithm based on a support vector machine (SVM), which was trained with a tenfold cross-validation.

The UK Biobank developed a QC algorithm to assess the quality of their own dataset only [11]. A classical machine learning approach was proposed to automatically identify problematic scans based on 190 image-derived features. These features are derived from both raw images as well as from derivatives after preprocessing. A Weka machine learning toolbox was used, with three separate classifier's outputs fused together, and a voting system combining the a posteriori probabilities of the three classifiers was used for the fusion. To train this QC algorithm, the quality of the first release of the Biobank MRI scans ($N=5816$) was assessed manually (number of raters unknown). For training, a stratified tenfold cross-validation was used. To test the algorithm, the second release of the Biobank was used, which was not manually labeled, and therefore, no performance measures could be derived.

Esteban et al. [12] developed MRIQC, a publicly available algorithm which extracts 64 image quality features and fits a binary classifier. The features are based on background evaluation of the raw MRI scan only. A supervised classification framework was used, composed of a random forest classifier (RFC), with a Leave-one-Site-out splitting for cross-validation, where a whole site is left out as a test set at each cross-validation fold. The quality of all MR volumes ($N=1102$) was first manually assessed by two raters, and they were given a label of "exclude," "doubtful," or "accept." For the binary classifier, pass consisted of the scans with an "accept" and "doubtful" label, whereas fail was composed of all manually "excluded" MRI scans.

Deep learning

Five studies utilized a deep learning approach [8, 19, 23, 25, 31], of three of which classify the MRI scans as either pass or fail [19, 23, 31] and the remaining two [8, 25] assess whether or not motion is present in the assessed scans.

The first article investigating the feasibility of automated detection and assessment of motion artifacts in MRI scans with a convolutional neural network (CNN) was published by Küstner and colleagues [25]. The input of the CNN was a patched image; it was also investigated which patch size was the best. For training and validation, MRI scans from 16 healthy volunteers were used. Each volunteer was scanned twice, and during the first acquisitions, the volunteers were instructed to hold their head still. During the second acquisition, volunteers were instructed to deliberately tilt their head side-to-side. For training and evaluation, a leave-one-out cross-validation approach was used. This CNN was able to localize and detect motion artifacts. Furthermore, it was shown that patch-based accuracy of detecting motion artifacts declined with decreasing patch size.

Fantini et al. [8] continued on the work of Küstner et al. [25], by including scans with more complex motion distortions ($N=203$). Furthermore, they also investigated the performance of four different networks, namely, Xception, InceptionV3, ResNet50, and Inception-Resnet, and applied transfer learning by pretraining the networks on the Imagenet dataset. One model was trained for each standard MR orientation (axial, coronal, and sagittal axes) on patches, and an artificial neural network was used to combine the outputs of the different networks to one output value. Also, a depth search was performed, attaching the binary classifier on distinct block output, and the best architecture depth was selected from the block layer that reported the best accuracy. A threefold cross-validation approach was used for the training of all architectures.

Sujit et al. [31] aimed to develop an algorithm that would evaluate the image quality of 3D T1w MRI scans using data from a large multicenter database ($N=1064$) which were classified by two raters. Their deep learning network was inspired by the VGG16 network, and one model was trained for each standard MR orientation (axial, coronal, and sagittal axes), and the output layer provided a slice-wise quality score. A second network consisting of a fully connected layer and an output layer was used to combine all the slice scores into one "volumetric" quality score. It was shown that this model provided good accuracy for classifying brain MRI scan quality.

Bottani and colleagues [19] developed an algorithm for automatic QC of MRI scans in a large clinical data warehouse ($N=3770$) which were rated by two raters. They aimed to discard scans which are not proper T1w brain MRI, identify scans with gadolinium, and recognize scans of bad, medium, and good quality. For the purpose of this paper, we will focus on the last two aims. MRI scans were classified into three tiers, i.e., good quality, medium quality, and bad quality. For the classification between bad vs medium/good and medium vs good, two separate networks were trained. It was trained using the cross entropy loss, which was weighted according to the proportion of scans per class for each task.

Another approach was used by Keshavan et al. [23] in which citizen scientists were used to visually QC MRI scans in order to acquire a large labeled dataset ($N=200$) which in turn can be used to train deep learning networks for automatic QC. A VGG16 network was pretrained on the Imagenet database, and further training and testing were done with the scoring of both the experts and the citizen scientists. Performance was reported by comparing the outcomes of this network and amplified training dataset with the outcomes of MRIQC [12].

Discussion

In this systematic review, we identified 18 unique algorithms used for quality control of structural T1w scans, of which ten use a rule-based approach, three use a classical machine learning approach, and five use a deep learning approach. The results of our systematic review revealed three key findings. Firstly, we identified a wide array of features incorporated within these algorithms, and even though there is little consistency across algorithms in terms of features, most of them utilize at least one feature assessing the entropy of the image. Secondly, the lack of consistent metrics and evaluation criteria hindered direct comparisons and highlighted the importance of establishing standardized performance measures within the field. Lastly, we observed significant variability in the selection of benchmarks employed during the development of QC algorithms across different approaches.

Features

By design, both rule-based and machine-learning algorithms use predefined features aimed at capturing image properties. A great number of features have been proposed, but only signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) have been used in three or more algorithms. The features can roughly be categorized into three categories: entropy features, image contrast features, and normative features. Initially, studies focused on entropy measures [20, 33] capturing the randomness in an image, and therefore, the majority of the features belong to this category because other studies build upon those. Features in more recent studies are categorized best as normative measures, which capture deviations of the image compared to the mean. Features extracted from the background of an image can be classified into either the image contrast or entropy category. These categories are commonly utilized in studies evaluating noise or blur, which tend to distribute evenly throughout the image, thus including the background. Features extracted from the foreground generally belong to the normative category, with the purpose of identifying deviations from the usual patterns. Features related to normative features capture a broader range of artifacts than the other two categories. All algorithms combine features from different categories to detect a wide range of artifacts, except those using the QEMDIM algorithm [21, 22]. In that case, the authors argue that a single feature, reflecting the distance from a benchmark, captures enough information to catch multiple sources for image degradation.

It should be noted that a minimal processing workflow is generally utilized to extract the different features, typically developed based on the MRI scan of healthy individuals without significant artifacts. As a result, the presence of artifacts and clinical deviations can potentially impact the performance of the workflow and thus the extracted features. In cases where a scan fails the processing workflow, it could be attributed to the presence of artifacts or clinical deviations such as the presence of a tumor. This creates a circularity in feature extraction, where features are used to quantify artifact presence, yet the extraction of features is influenced by the presence of irregularities.

Deep learning algorithms do not use predefined features, and by design, the factors that contribute to the classification by these networks are not always apparent, which limits interpretability. In the included articles, different architectures of CNNs are being used. Both traditional sequential network [23, 31] and network-in-network architectures are being used for QC [8]. Fantini et al. [8] showed that the different architectures lead to different results but only compared network-in-network architectures with each other. The different architectures were trained on the same dataset, and thus, a fair comparison on performance could be made. Ultimately, deep learning algorithms can only be compared with each other, or other types of algorithms, based on the final performance measures.

Performance

One consistent finding is that the use of dataset-specific optimization creates a circular process that inflates performance measures. During the training of the machine-learning and deep learning algorithms, or when setting the thresholds for the rule-based algorithms, an iterative approach is used in all studies that repeatedly refers back to the dataset for fine-tuning. This suggested that the performance of the algorithms was artificially inflated on that particular dataset. Consequently, comparing the performance of two or more algorithms becomes challenging, as the performance measures only reflect performance on specific datasets. The limitations of dataset-specific optimization become apparent when independent validation is conducted, causing the accuracy to decline by approximately 11%. Noteworthy, no independent validation of the thresholds in rule-based algorithms was performed, except for the ghosting threshold in [20]. Independent validation is performed in one classical machine-learning

algorithm [12] and three deep learning algorithms [8, 19, 31]. As can be seen in Table 1, various performance measures have been used. Often, accuracy was chosen to evaluate the model, since it coincides well with the general aim of developing a QC algorithm, i.e., to predict the class of unseen MRI scan accurately. However, it might not be the best performance measure in all cases, as it conveys well only when all classes (pass or fail) have similar prevalence in the data. However, this is not the case, as the majority of the studies utilizing visual QC have less than 40% of the used dataset categorized as fail. Additional measures like sensitivity, specificity, PPV, and NPV are therefore needed to provide a more complete overview of the performance.

Datasets

Specific characteristics of individual datasets introduce significant variations and pose challenges when comparing algorithms. The majority of algorithms are trained on datasets incorporating participants with pathologies, especially Alzheimer's disease (ADNI) and autism spectrum disorder (ABIDE), while some contain solely healthy individuals. The choice to include scans where pathologies might be present seems to be driven by the fact that those types of datasets are relatively big and publicly available. The studies including their own datasets often include solely healthy individuals and are notably smaller. There is a risk that algorithms trained on healthy datasets may misclassify pathologies as artifacts [34], although they may exhibit higher sensitivity to small artifacts. Variability in datasets is also found in the age distribution of the included participants. For two algorithms, the age distribution is not reported [20, 28], but in the majority of utilized datasets, adults (age range 18–65) were included. However, four datasets included children (age < 18) only, four included adults and elderly (age 18 +), and two datasets were a combination of children and adults (age < 65). It is shown that there are differences in cortical gray matter as a function of age between children and adults [35], as well as that the cortical thickness shows regional and temporal specificity with development [36]. Differences are also found between adults and elderly, like atrophy or the expansion of the ventricular system [37]. Therefore, algorithms developed using datasets exclusively containing either children or adults may restrict their generalizability across diverse age ranges as the variability may be seen as image degradation [38]. The generalizability of algorithms can also be limited by the choice of T1w sequence employed in their training. Algorithms trained on 2D scans may not be applicable to 3D scans, and the specific sequence utilized could also have an impact, potentially leading to variations in artifacts.

Benchmark

To evaluate algorithm performance, a benchmark was established for all algorithms. Either visual QC or synthetically degraded images have been used for this goal. For visual QC, a predefined protocol is used to mitigate subjectivity [39], as there are inherent variations in determining an acceptable level of image quality [5]. In the case of synthetically degraded images, filters are applied to simulate various types of artifacts. For reproducibility purposes, the details of the filters used and the corresponding parameters employed should be specified. However, some studies lack information on their protocols, such as the number of raters and consensus methods. In none of the studies using synthetically degraded images, a rationale was provided for filter selection or parameter scaling.

To evaluate the visual QC output, most studies [8, 11, 23, 26, 30–32] use a binary pass or fail classification, while two studies [12, 24] introduce a "doubtful" category, which is later merged with the pass category. This decision is based on achieving higher agreement between automatic QC and visual QC. However, it might be more advantageous to merge the "doubtful" category with the fail category, rather than the pass category, allowing users to focus their visual QC solely on the failed scans in case of semi-automatic QC [11, 24, 39]. Synthetically degraded images, on the other hand, are measured on a continuous scale, often through percentages or scaling of the filter parameters. However, it remains unclear how this scaling relates to real-world artifacts, raising questions about the transferability and practical interpretation of these synthetic measures.

Both methods are limited in that there is a lack of consistency in the benchmark, and thus, direct comparisons and generalizations of the algorithms are hindered. In the case of the visual QC, there have been attempts to develop uniform and robust visual QC rating systems to improve replication and comparability between studies [39]. In the case of synthetically degraded images, there are concerns about the generalizability of the findings to real-world clinical settings, where the types and severity of artifacts may differ significantly from those artificially induced. The lack of consistency in the benchmark is also due to the intended use of the scans.

Level of quality control

Benchmarks are established based on the intended data usage. The extent of QC may vary depending on whether the data are intended for clinical diagnosis or for research studies. In clinical diagnosis, T1w MRI scans are primarily used for individual patient assessment, so the QC process tends to focus on ensuring that gross abnormalities and readily noticeable

anomalies are detectable. Therefore, the QC requirements for clinical diagnosis may be somewhat more flexible regarding minor imperfections in scans, as long as they do not impede the identification of obvious clinical issues. In research studies involving large group comparisons or detailed quantification, the objective often extends beyond the mere detection of gross anomalies, and therefore, the QC process may demand a higher level of data quality. The QC procedures in this case are geared towards ensuring the reliability of quantitative data and may be less tolerant of variations in data quality.

Future directions

To evaluate and compare algorithms in a non-biased manner, the algorithms should be trained on the same dataset. This way, it is ensured that comparisons between algorithms are focused on the inherent capabilities and limitations of the algorithms themselves rather than being influenced by specific characteristics of individual datasets. By encompassing a wide range of ages, pathologies, and artifact prevalence, the generalizability and performance across different scenarios might improve. Additionally, a consensus [8] should be reached, for which the development of standardized protocols for visual QC or synthetically degraded images could be beneficial. Also, it might be useful to work with a benchmark composed of both real-world scans that are visually checked and synthetically degraded images, since this would reduce the subjectivity but also increase the generalizability of findings to real-world clinical settings. To accommodate variations in the required level of QC, multiple thresholds can be employed, or a set of trained classifiers can be utilized. This approach enables customization of QC procedures to align with the specific needs and intended usage of the data.

Furthermore, researchers continue to explore new features to improve the algorithms, aiming for more robust and discriminative measures. However, it is equally important to investigate the discriminative power of the individual features that are already being used, particularly for specific artifacts. Additionally, explainable deep learning could be used to gain insight into the decision-making process of the deep learning-based algorithms and, thus, the image characteristics such algorithms focus on. Understanding which feature or set of features is most effective in capturing specific artifacts can provide valuable insights for further refinement or the algorithms.

Additionally, generative models have recently attracted much attention in quality control or anomaly detection, due to their unique ability to generate new data when there is a lack of data that represents the anomalous behavior and the ability to apply representation learning [40]. This might also be useful for the QC of T1w MRI scans.

Finally, the impact of automatic QC on daily clinical practice can be significant. It can streamline the process of reviewing and verifying the quality of the scan, therefore saving time for healthcare professionals, as it can reduce the need for manual, time-consuming assessments. Also, it ensures a standardized and consistent approach for evaluation. Furthermore, if the automatic QC can be performed in real time when the patient is in the scanner, it allows for timely rescans or adjustments, preventing the need for patients to return for additional imaging sessions, and additionally reducing costs for the hospitals.

Conclusion

To the best of the authors' knowledge, this is the first review on (semi-)automatic QC algorithms for T1w MRI scans. The detected algorithms employ diverse approaches and features, with an emphasis on entropy measures. However, comparing algorithm performance was challenging due to dataset-specific optimization, inflating results and hindering cross-dataset comparisons. Also, variability and missing information in the benchmark were found, including unclearities in protocols and limited information on filter selection and parameter scaling. Despite these limitations, our review provides valuable insights into the landscape of QC algorithms for structural T1w scans. The implications of these findings call for future research and collaboration to establish guidelines and best practices, to ultimately enhance the reliability and effectiveness of QC algorithms in this domain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00234-023-03256-0.

Funding The collaboration project (Amsterdam UMC 2011227) is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Declarations

Competing interests Richard Joules reports a relationship with IXICO Technologies Ltd that includes employment. Oscar Pena-Nogales reports a relationship with QMENTA Inc. that includes employment. Paulo Reis Rodrigues reports that financial support was provided by QMENTA Inc. Paulo Reis Rodrigues reports a relationship with QMENTA Inc. that includes board membership, employment, and equity or stocks. Paulo Reis Rodrigues has patent #WO2020210826A1 pending to Mint Labs Inc d/b/a QMENTA Inc. Robin Wolz reports that financial support was provided by IXICO Technologies Ltd. Robin Wolz reports a relationship with IXICO Technologies Ltd that includes employment and equity or stocks.

Informed consent For this type of study, formal consent is not required.

References

1. Shaikh F, Dupont-Roettger D, Dehmeshki J, Awan O, Kubassova O, Bisdas S (2020) The role of imaging biomarkers derived from advanced imaging and radiomics in the management of brain tumors. Front Oncol 10:559946
2. Oishi K, Faria AV, Yoshida S, Chang L, Mori S (2013) Quantitative evaluation of brain development using anatomical MRI and diffusion tensor imaging. Int J Dev Neurosci 31(7):512–524
3. McEvoy LK, Brewer JB (2010) Quantitative structural MRI for early detection of Alzheimer's disease. Expert Rev Neurother 10(11):1675–1688
4. Paolillo A, Coles AJ, Molyneux PD, Gawne-Cain M, MacManus D, Barker GJ, Miller DH (1999) Quantitative MRI in patients with secondary progressive MS treated with monoclonal antibody Campath 1H. Neurology 53(4):751–7
5. Scheltens P, Launer LJ, Barkhof F, Weinstein HC, van Gool WA (1995) Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability. J Neurol 242(9):557–560
6. Gardner EA, Ellis JH, Hyde RJ, Aisen AM, Quint DJ, Carson PL (1995) Detection of degradation of magnetic resonance (MR) images: comparison of an automated MR image-quality analysis system with trained human observers. Acad Radiol 2(4):277–281
7. Kruggel F, Turner J, Muftuler LT, and I. Alzheimer's disease neuroimaging (2010) Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. Neuroimage 49(3):2123–33
8. Fantini I, Yasuda C, Bento M, Rittner L, Cendes F, Lotufo R (2021) Automatic MR image quality evaluation using a Deep CNN: a reference-free method to rate motion artifacts in neuroimaging. Comput Med Imaging Graph 90:101897
9. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Collins R (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):e1001779
10. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, W.U.-M.H. Consortium (2012) The human connectome project: a data acquisition perspective. Neuroimage 62(4):2222–31
11. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, Smith SM (2018) Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166:400–424
12. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ (2017) MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One 12(9):e0184661
13. Keshavan A, Datta E, M.M. I, Madan CR, Jordan K, Henry RG (2018) Mindcontrol: a web application for brain segmentation quality control. Neuroimage 170:365–372
14. Klapwijk ET, van de Kamp F, van der Meulen M, Peters S, Wierenga LM (2019) Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. Neuroimage 189:116–129
15. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Moher D (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 62(10):e1-34
16. Moher D, Liberati A, Tetzlaff J, Altman DG, P. Group (2010) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 8(5):336–41
17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, P.-P. Group (2015) Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 4(1):1
18. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, . . . Groupdagger P (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 170(1):51–58
19. Bottani S, Burgos N, Maire A, Wild A, Stroer S, Dormont D, A.S. Group (2022) Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Med Image Anal 75:102219
20. Gedamu EL, Collins DL, Arnold DL (2008) Automated quality control of brain MR images. J Magn Reson Imaging 28(2):308–319
21. Ikushima Y, Tokurei S, Tarewaki H, Morishita J, Yabuuchi H (2022) A novel algorithm for comprehensive quality assessment of clinical magnetic resonance images based on natural scene statistics in spatial domain. Magn Reson Imaging 92:203–211
22. Jang J, Bang K, Jang H, Hwang D, and I. Alzheimer's disease neuroimaging (2018) quality evaluation of no-reference MR images using multidirectional filters and image statistics. Magn Reson Med 80(3):914–924
23. Keshavan A, Yeatman JD, Rokem A (2019) Combining citizen science and deep learning to amplify expertise in neuroimaging. Front Neuroinform 13:29
24. Kim H, Irimia A, Hobel SM, Pogosyan M, Tang H, Petrosyan P, Toga AW (2019) The LONI QC system: a semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data. Front Neuroinform 13:60
25. Küstner T, Liebgott A, Mauch L, Martirosian P, Bamberg F, Nikolaou K, Gatidis S (2018) Automated reference-free detection of motion artifacts in magnetic resonance images. MAGMA 31(2):243–256
26. Mortamet B, Bernstein MA, Jack CR Jr, Gunter JL, Ward C, Britson PJ, . . . I. Alzheimer's disease neuroimaging (2009) Automatic quality assessment in structural brain magnetic resonance imaging. Magn Reson Med 62(2):365–72
27. Osadebey M, Pedersen M, Arnold D, Wendel-Mitoraj K (2017) No-reference quality measure in brain MRI images using binary operations, texture and set analysis. IET Image Proc 11(9):672–684
28. Osadebey ME, Pedersen M, Arnold D, Wendel-Mitoraj K and A.s.D.N. Initi (2017) The spatial statistics of structural magnetic resonance images: application to post-acquisition quality assessment of brain MRI images. Imaging Sci J 65(8):468–483
29. Osadebey ME, Pedersen M, Arnold DL, Wendel-Mitoraj KE (2018) Blind blur assessment of MRI images using parallel multiscale difference of Gaussian filters. Biomed Eng Online 17(1):76
30. Pizarro RA, Cheng X, Barnett A, Lemaitre H, Verchinski BA, Goldman AL, Mattay VS (2016) Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm. Front Neuroinform 10:52
31. Sujit SJ, Coronado I, Kamali A, Narayana PA, Gabr RE (2019) Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. J Magn Reson Imaging 50(4):1260–1267
32. White T, Jansen PR, Muetzel RL, Sudre G, El Marroun H, Tiemeier H, Verhulst FC (2018) Automated quality assessment of structural magnetic resonance images in children: comparison

with visual inspection and surface-based reconstruction. Hum Brain Mapp 39(3):1218–1231
33. Woodard JP, Carley-Spencer MP (2006) No-reference image quality metrics for structural MRI. Neuroinformatics 4(3):243–262
34. Krupa K, Bekiesinska-Figatowska M (2015) Artifacts in magnetic resonance imaging. Pol J Radiol 80:93–106
35. Jernigan TL, Tallal P (1990) Late childhood changes in brain morphology observable with MRI. Dev Med Child Neurol 32(5):379–385
36. Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW (2003) Mapping cortical change across the human life span. Nat Neurosci 6(3):309–315
37. Fjell AM, Walhovd KB (2010) Structural brain changes in aging: courses, causes and cognitive consequences. Rev Neurosci 21(3):187–221
38. Vogelbacher C, Mobius TWD, Sommer J, Schuster V, Dannlowski U, Kircher T, Bopp MHA (2018) The Marburg-Munster Affective Disorders Cohort Study (MACS): a quality assurance protocol for MR neuroimaging data. Neuroimage 172:450–460
39. Backhausen LL, Herting MM, Buse J, Roessner V, Smolka MN, Vetter NC (2016) Quality control of structural MRI images applied using freesurfer-a hands-on workflow to rate motion artifacts. Front Neurosci 10:558
40. Sabuhi M, Zhou M, Bezemer CP, Musilek P (2021) Applications of generative adversarial networks in anomaly detection: a systematic literature review. Ieee Access 9:161003–161029