

Original Research Articles

Automatic detection of simulated artifacts on T1w magnetic resonance images: Comparing performance of different QC strategies

Janine Hendriks, MSc^{1,2}^a, Michelle G. Jansen, PhD^{1,2,3}, Richard Joules, PhD⁴, Óscar Peña-Nogales, PhD⁵, Paulo R. Rodrigues, PhD⁵, Frederik Barkhof, PhD^{1,6}, Anouk Schranter, PhD^{2,3}, Henk J.M.M. Mutsaerts, PhD^{1,2}

¹ Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Location VUmc, Amsterdam, The Netherlands,

² Amsterdam Neuroscience, Amsterdam, The Netherlands, ³ Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Location AMC, Amsterdam, The Netherlands, ⁴ Ixico, London, UK, ⁵ QMENTA Inc., Boston, MA, USA, ⁶ Queen Square Institute of Neurology and Centre for Medical Image Computing, University College London, London, UK

Keywords: T1w, quality control

<https://doi.org/10.52294/001c.172318>

Aperture Neuro

Vol. 6, 2026

The reliability of MRI-derived measures critically depends on image quality. Poor-quality scans can obscure anatomical detail and compromise the accuracy of automated image analysis, underscoring the need for robust quality control (QC) procedures. Automated QC offers scalability for large neuroimaging datasets, yet the comparative performance of different approaches for detecting specific artifact types remains poorly understood.

We systematically compared rule-based (RB), classical machine learning (ML), and deep learning (DL) QC algorithms using 1,000 high-quality T1w scans. Four artifact types, blurring, ghosting, motion, and noise, were synthetically introduced across ten severity levels using TorchIO, yielding 40,000 degraded images. Visual QC of a subset confirmed strong inter-rater reliability (Krippendorff's $\alpha = 0.82$, mean Spearman's $\rho = 0.87$). RB and ML models used 62 image quality metrics from MRIQC, whereas DL models were trained directly on minimally preprocessed images. Models were trained with participant-level five-fold cross-validation and tested on an independent dataset.

DL models achieved the highest overall performance across simulated artifact types (Youden's Index [YI] = 0.83–0.97). RB and ML performed comparably at high artifact severities (YI ≥ 0.75) but showed limited sensitivity to subtle simulated ghosting and noise (YI ≤ 0.15). Feature analysis indicated that RB relied primarily on normative metrics, whereas ML flexibly adapted feature use by artifact type and severity.

These findings highlight DL's superior generalizability for detecting subtle simulated artifacts and provide a proof-of-concept for selecting QC strategies for similar benchmarking tasks.

1. INTRODUCTION

Magnetic resonance imaging (MRI) is a powerful, non-invasive diagnostic tool that provides detailed images of brain anatomy, making it invaluable for investigating neurological conditions in both research and clinical settings. T1-weighted (T1w) images are frequently used to derive conventional measures of brain health, such as gray matter volume or cortical thickness. However, the reliability of MRI derivatives is heavily dependent on the quality of the ac-

quired MRI scans. Poor image quality not only makes it challenging for radiologists to discern relevant anatomical detail but may also compromise the accuracy and precision of automated image analysis algorithms.^{1–4} Therefore, robust quality control (QC) measures are essential to identify and appropriately manage poor-quality images that could negatively impact statistical power in research studies. Beyond guiding exclusion decisions, QC metrics can also be used to modulate confidence in results, serving as covariates or weights in subsequent analyses.^{5,6}

^a Corresponding Author:

Janine Hendriks, MSc (j.m.hendriks@amsterdamumc.nl)

Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Location VUmc
Amsterdam, The Netherlands

Traditional QC methods typically involve visual inspection by radiographers, radiologists, or researchers, which is both time-consuming and subjective.^{7,8} In contrast, automated QC could offer significant advantages, such as improved reproducibility and standardization by eliminating human bias, and increased efficiency because of computer speed and parallelization potential. These advantages make automated QC a feasible solution for large research studies or potentially even real-time feedback on the scanner in the future. Despite these benefits, the implementation of automated QC has yet to be standardized across institutions, in part due to the diversity of research protocols, scanner hardware, and study aims. Research groups often develop their own QC algorithms to assess image quality, particularly focusing on artifact detection before further post-processing.⁹⁻¹¹ Some of these algorithms are fully automated in performing the assessment (automatic QC), while others still require human intervention at key points (semi-automatic QC). As the field advances, establishing consensus on systematic QC approaches is essential for enhancing the reliability of studies and improving comparability across different research centers and clinical settings.

Semi-automatic and automatic QC algorithms can be categorized into three categories: rule-based (RB), classical machine learning (ML), and deep learning (DL) approaches.¹¹ RB approaches establish thresholds for predefined image quality features, such as signal-to-noise ratio (SNR), to flag scans that fall outside these thresholds.¹² ML algorithms rely on similar predefined features but involve the training of classifiers to differentiate image quality levels.¹⁰ DL algorithms differ from RB and ML as DL does not rely on predefined quality features but instead uses all available image information to classify the images.¹³⁻¹⁵

Algorithms from each of these categories have demonstrated high accuracy within (mostly) single-cohort studies, which makes it challenging to compare QC studies due to variations in MRI scanner hardware, MRI protocols, and patient populations.¹⁶ The variability introduced by these factors can lead to discrepancies in image characteristics that are not accounted for by the algorithms trained on a specific dataset. Moreover, the relative performance of these methods is difficult to quantify, as reported accuracy is often contingent on the specific context of the studies and local practices, rather than on universally applicable quality criteria.¹¹ The lack of standardized benchmarking datasets and evaluation metrics further complicates the comparison of different QC methods. For example, some studies may emphasize sensitivity to certain artifacts over others, such as blurring, noise, head motion, or ghosting, which can bias the reported performance metrics.¹⁷

A main challenge in automatically detecting (subtle) artifacts lies in the fact that only a small subsection of MRI scans contains severe artifacts, which are often not shared publicly. On the other hand, most MR images may contain subtle artifacts, which are not detected visually but may already affect statistical power. Simulated artifacts may help in understanding the effects of subtle artifacts, although these are restricted in real-world generalizability.

As a proof of concept, the present study compares the ability of automatic QC algorithms to detect simulated artifacts. The present study focused on a single, standardized T1w MPRAGE protocol from the Alzheimer's Disease Neuroimaging Initiative (ADNI), to provide a controlled benchmark for comparing the performance of automatic QC algorithms under known synthetic artifact types and severities, to better understand their strengths and limitations. We 1) evaluate whether different severity levels of these synthetic artifacts can be detected visually by human readers; 2) compare the artifact detection performance between RB, ML, and DL; and 3) investigate whether image quality metric (IQM) feature importance differs between RB and ML.

2. METHODS

2.1. DATA INCLUSION

Data for this study were drawn from the ADNI dataset, a multisite study of older adults with normal cognition, mild cognitive impairment, or Alzheimer's disease.¹⁸ Acquisition and reconstruction details for the T1w MRI data were obtained from the ADNI MRI Scanner Protocols (<https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/mri-scanner-protocols/>) and MRI Pre-processing documentation (<https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/mri-pre-processing/>), together with scan-level metadata in the ADNI Image & Data Archive.

We selected high-quality 3D T1w MRI scans that are free of apparent artifacts, using visual QC scores from three independent groups. From the full pool of 17,141 3D T1w MRI scans¹⁹ (downloaded August 2024), 1,806 3T images had three visual QC scores available for the unprocessed T1w scans, one score from each of the three independent sources: the Dementia Research Centre of University College London (UCL), the NeuroImaging & Surgical Technologies (NIST) Lab of McGill University, and the Public Open Neuroimaging Data Reduction for Artificial Intelligence (PondrAI) QC. The UCL laboratory performed visual QC using a 3-level scale: 1) "ok", 2) "borderline", and 3) "fail". Their assessment focused on identifying the presence of artifacts such as blurring, ghosting, motion, noise, and ringing.²⁰ The McGill laboratory involved 12 raters who performed visual QC using QRater with a 3-level scale: 1) "ok", 2) "warning", and 3) "fail". The QRater visual QC protocol outlines criteria based on coverage, intensity uniformity, motion, and noise.²¹ Finally, PondrAI QC is a public repository containing visual QC of a variety of public datasets. The images were rated by a single rater on a 6-level scale, with 1 indicating good image quality and 6 indicating poor image quality, based on contrast, coverage, noise, and ringing.²²

We rescaled all ratings to a 3-point scale (1–3), with 1 indicating good image quality and 3 indicating poor image quality and summed them to a composite QC score (3–9). A total of 1,462 scans received a composite QC score of 3. Using a fixed random seed, we pseudorandomly selected a subset of 1,000 scans for subsequent analysis (for acquisi-

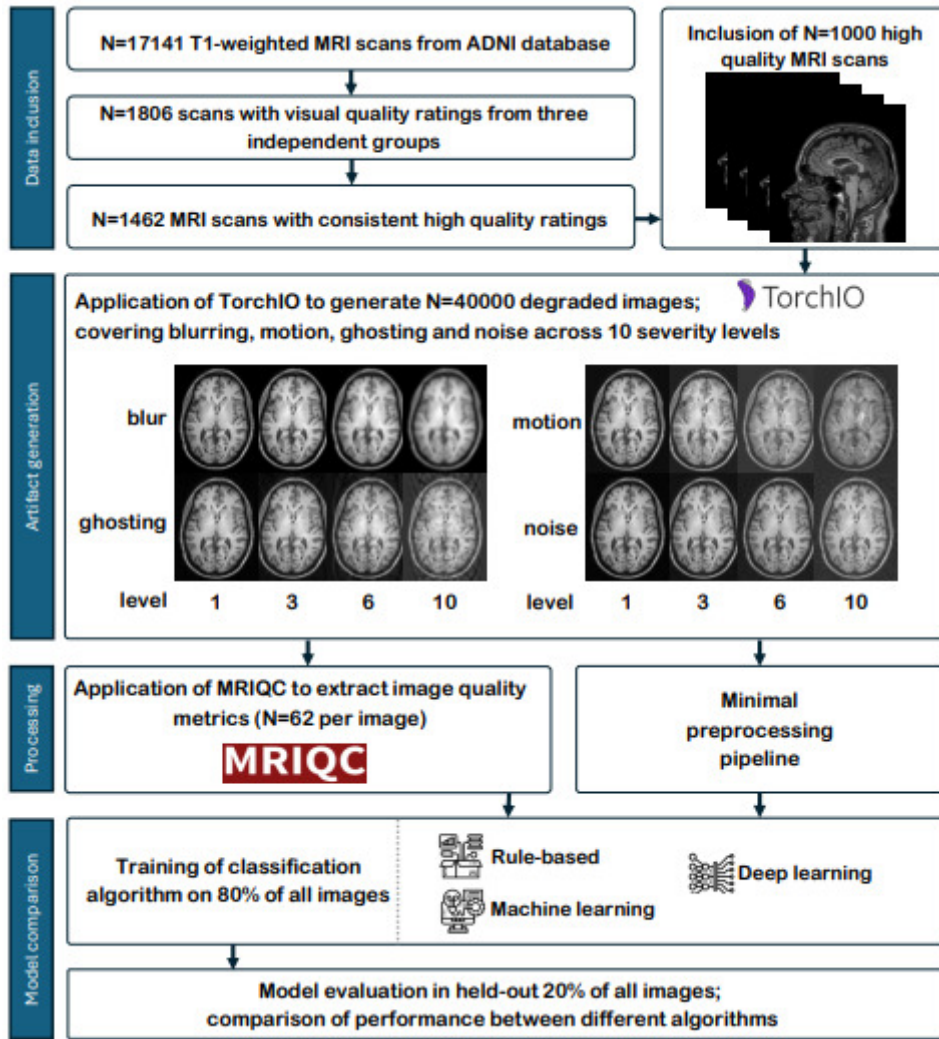


Figure 1. Schematic overview of methods.

tion details, see [Table 1](#)). These 1,000 MRI scans included data from 694 unique participants; 445 participants had only one session, 200 had two sessions, 41 had three sessions, and 8 had four sessions.

2.2. ARTIFACT GENERATION

We added four types of simulated artifacts (blurring, ghosting, motion, and noise) to the 1,000 MRI scans using TorchIO.²³ As TorchIO’s artifact simulations are dependent on image intensities, we normalized the mean and variance of each scan prior to applying the simulations. These normalized images will be referred to as non-degraded images. For each artifact type, we generated ten severity levels ([Figure 1](#)) by linearly adjusting the relevant parameters of the TorchIO functions. The lower and upper boundaries for each artifact’s parameter range (sections 2.1–2.4) were defined based on previous work,²⁴ ensuring a clinically realistic spectrum. This approach resulted in 1,000 MRI scans $\times$ 4 artifact types $\times$ 10 severity levels = 40,000 degraded images, in addition to the 1,000 non-degraded images.

2.2.1. BLURRING

TorchIO’s *RandomBlur* function applies a Gaussian kernel to blur the image along each axis.²³ By linearly increasing the standard deviations of the kernels (0.2 to 2), we created various levels of blurring artifact severity.

2.2.2. GHOSTING

TorchIO’s *RandomGhosting* function generates ghosting artifacts by Fourier transforming the image to k-space, creating several shifted copies of the k-space data along a specified axis, scaling these copies by a ghosting intensity parameter, and combining them with the original before inverse Fourier transforming back to the original space. We set the axis to anterior-posterior and linearly increased the ghosting intensity (0.15 to 1.5) for ten severity levels.

2.2.3. MOTION

TorchIO’s *RandomMotion* function simulates rigid-body motion by applying a sequence of random rotations and translations in k-space.^{23,25} The interpolation method was set to b-spline, and by linearly increasing both rotation (1

to 10°) and translation parameters (1 to 10 mm) in tandem, we created ten severity levels.

2.2.4. NOISE

TorchIO's *RandomNoise* function adds zero-mean Gaussian noise without spatial correlation.²⁶ By linearly increasing the noise standard deviations (0.03 to 0.3), we created ten severity levels. If this resulted in voxel intensities below 0, these were retained (no intensity clipping was performed).

2.3. VISUAL QC OF DEGRADED SCANS

To evaluate how the simulated artifacts relate to those typically observed in real-world scans, we conducted visual QC on a subset of the data using the PondrAI QC rating scale (1 = perfect quality, 6 = terrible quality).²² Three thousand images were randomly selected from the 40,000 degraded images by shuffling the eligible image pool and then selecting the first 3,000, ensuring a representative distribution across all artifact types and severity levels (roughly 75 scans $\times$ 4 artifacts $\times$ 10 severity levels). Prior to rating, all nine raters completed a standardized training protocol by reviewing the manual provided by PondrAI²² and practicing on a separate set of 100 images randomly selected from the remaining pool (excluding the 3,000 chosen images). A consensus meeting followed the practice session to align interpretations of the rating criteria. Each rater was assigned a subset of 1,000 images such that each image was rated by three raters.

We assessed inter-rater reliability using Krippendorff's α and the average pairwise Spearman's rank correlation coefficient (ρ) using R (version 4.4.1; R Foundation for Statistical Computing, Vienna, Austria). Krippendorff's α was chosen because it is suitable for ordinal data, accommodates missing values, and adjusts for agreement expected by chance, where agreement is indicated as ≥ 0.8 high, 0.67–0.79 moderate and < 0.67 poor.²⁷ Spearman's ρ was used to quantify the consistency of the relative ranking of images between raters, with correlations indicated as < 0.39 weak, 0.40–0.59 moderate, 0.60–0.79 strong, and ≥ 0.8 very strong.

Krippendorff's α was computed using the “ordinal” method, with 95% confidence intervals (CIs) estimated via bootstrapping (1,000 iterations). For Spearman's ρ , all pairwise correlations between raters were calculated for each artifact type using the “*pairwise.complete.obs*” option to handle missing data, and then averaged.²⁸ Confidence intervals [CI] for Spearman's ρ were calculated using the Jackknife Euclidean likelihood-based inference approach.²⁹

2.4. MODEL TRAINING

We evaluated three different types of artifact classification algorithms: RB, ML, and DL. For each artifact type and severity level, a separate model was trained to perform binary classification (e.g., the presence vs absence of blurring at severity level 5), resulting in 40 RB models, 40 ML models, and 40 DL models. No combinations of artifacts or

severities were trained. All algorithms were implemented using scikit-learn (version 1.2.2).

To prevent data leakage, the data were split at the participant level, with all longitudinal scans from each participant (section 1) assigned exclusively to one set, resulting in a training set of 765 images and a test set of 235 images. Within the training set, 5-fold cross-validation was performed at the participant level: in each fold, 612 images (80% of the training set) were used for model training, and 153 images (20% of the training set) were used for validation, again ensuring that all data from a given participant remained within a single set. For each artifact type and severity level, separate datasets were created, resulting in a total of 40 datasets (4 artifact types $\times$ 10 severity levels). All three models were trained independently for each artifact and each severity level, resulting in 120 models.

2.4.1. INPUT FEATURES RB- AND ML-BASED CLASSIFICATION MODELS

As input features for the RB and ML-based classification algorithms, we derived IQMs from the 1,000 non-degraded and 40,000 degraded images using MRIQC (version 22.0.1). Briefly, MRIQC performs head motion estimation with AFNI (version 22.0.17),³⁰ followed by brain extraction with SynthStrip (version 1),³¹ tissue segmentation with FSL FAST (version 5.0.11),³² and several image registration steps to MNI space with ANTS (version 2.3.3.dev168-g29bdf),³³ without accounting for dependencies between multiple sessions from the same participant. Subsequently, MRIQC derives a set of 68 IQMs based on the Quality Assessment Protocol.³⁴ The six spatial resolution IQMs (x/y/z matrix size and voxel-size) were excluded because the spatial resolution does not depend on artifacts, resulting in 62 IQMs (Supplementary Table 1).

To distinguish normal from abnormal IQM values, we reuse predefined quality thresholds if they existed in previous literature, which was the case for five IQMs. These reference-based IQMs have a single absolute quality threshold (e.g., intracranial tissue volume fraction).¹⁰ The other 57 non-reference-based IQMs lack established thresholds and are typically used for relative quality comparisons (e.g., SNR).¹⁰ To harmonize vendor-based batch effects among the non-reference IQMs, we calculated the mean and standard deviation for each IQM per vendor in the 1,000 non-degraded images and used these to calculate IQM Z-scores for the 40,000 degraded images.

2.4.2. RB CLASSIFICATION MODEL

To develop and evaluate the RB classification model, we applied a multi-stage approach within each cross-validation fold, with the primary goal of determining optimal thresholds at the individual IQM level and then deriving a single threshold to classify the overall image quality. Thresholds for reference-based IQMs were fixed based on previously established values in MRIQC.¹⁰ The non-reference-based IQM thresholds were determined within each training fold by evaluating a range of z-score thresholds for each non-reference IQM (from $z = -5$ to 5, in increments of 0.05) and

selecting the threshold that maximized the Youden Index (YI; [specificity + sensitivity – 1]) in the validation set.¹²

Subsequently, the reference-based IQM thresholds and optimized non-reference-based IQM thresholds were used to binarize IQMs in the training and validation subsets of that fold. Per image, the binary IQM values (1 = good quality; 0 = poor quality) were summed to a compound quality score, ranging from 0 (all IQMs flagged as poor quality) to 62 (all IQMs flagged as good quality). To define the final classification threshold for artifact presence, we used the compound scores from the non-degraded images within the training subset of that fold as a reference and set the cut-off as the mean + 1 standard deviation.

2.4.3. ML-BASED CLASSIFICATION MODEL

The ML-based classification algorithm used nested cross-validation to train and evaluate a random forest classifier (RFC).¹⁰ Model training was performed using five outer folds for feature selection and model evaluation, and five inner folds for hyperparameter tuning.

To focus on the most informative IQMs while minimizing spurious or site-related influences, a two-step feature selection pipeline was applied in each outer fold.¹⁰ First, an extra trees classifier was trained to predict site using all IQMs,³⁵ and the IQMs most predictive of site were removed. Second, a Winnow-based selection was performed to retain only those features that provided more information than a randomly generated feature. Subsequently, this reduced feature set was used to optimize the inner folds hyperparameters.

Within each inner fold, a grid search was conducted to optimize the RFC's hyperparameters for YI (number of estimators [10, 100, 500], maximum features [2, 4, 6, 8, 10, 15, 20], maximum depth [10, 20, 30, 50], and minimum samples split [5, 10, 20]).

2.4.4. DL-BASED CLASSIFICATION MODEL

Unlike RB and ML-based classification models, which rely on IQMs extracted with MRIQC, the DL classification algorithms were directly trained on minimally preprocessed non-skull-stripped images. To keep the preprocessing comparable between the classification algorithms, we modified an existing preprocessing pipeline¹⁵ to correspond with the preprocessing of MRIQC. Bias field correction was applied using the N4ITK method,³⁶ and spatial registration to MNI space was performed using ANTs. The minimum and maximum MNI space T1w image values were rescaled to a range $\in [0,1]$.

We trained an existing convolutional neural network (CNN) architecture with its original hyperparameters and loss functions,¹³ using five convolutional blocks followed by three fully connected layers. Each convolutional block consisted of a convolutional layer, a batch normalization layer, a rectified linear unit, and a max pooling layer. This CNN previously demonstrated the best performance for MRI QC in a large heterogeneous clinical dataset.

2.5. FINAL MODEL EVALUATION

For all 120 models, the configuration (e.g., thresholds, hyperparameters, model weights) achieving the highest YI across all five folds was selected. Notably, for ML-based classification, this involved the outer folds of the nested cross-validation procedure. Subsequently, the final model performance of each model was evaluated on the independent test set ($n = 235$). In addition to our primary performance metric YI, we also calculated accuracy, balanced accuracy, F1 score, precision, and recall to evaluate prediction accuracy.

2.6. FEATURE IMPORTANCE OF RB AND ML MODELS

To explore model interpretability of RB and ML algorithms, IQM importance was examined across the artifact types and severity levels. For each of the 80 RB and ML models, the most important IQMs were identified as those with the highest YI in RB models and those retained after feature selection in ML models, and these were recorded across all artifact and severity combinations for the final model.

For interpretation purposes, IQMs are grouped into four main categories: entropy, contrast, inhomogeneity, and normative IQMs.^{9-11,34} Entropy IQMs, such as the entropy focus criterion (EFC), capture the randomness or complexity in image intensity distributions. Image contrast IQMs, for example the contrast-to-noise ratio (CNR), quantify the distinguishability between different tissue types or structures. Intensity inhomogeneity IQMs, such as intensity non-uniformity (INU), assess the low frequency spatial uniformity of image intensity. Normative IQMs, including metrics like the overlap of tissue probability maps and the scan segmentations (TPM overlap), compare image characteristics to a reference population or template to detect deviations from expected norms. For each artifact type, Spearman's rank correlations (ρ) were computed between severity levels (ordinal, 1–10) and IQM values (EFC, CNR, INU, TPM). ρ values were interpreted according to the following bins: negligible ($< |0.10|$), weak ($|0.10-0.39|$), moderate ($|0.40-0.69|$), strong ($|0.70-0.89|$), and very strong ($> |0.90|$).³⁷

3. RESULTS

3.1. DATASET CHARACTERISTICS

Demographic characteristics of the study population, including age, sex, and scanner vendor distribution for the 1,000 non-degraded images are summarized in [Table 1](#). Descriptive statistics of the 62 IQMs of these images, stratified by vendor, are provided in Supplementary Table 1.

3.2. VISUAL QC RATINGS ACROSS SIMULATED ARTIFACTS

Visual QC ratings increased with artifact severity for most artifact types ([Figure 2](#)). Overall, higher simulated artifact severity was associated with higher QC ratings. *Blurring* and *ghosting* showed a trend of linear increase across sever-

Table 1. Demographic and scanner-related characteristics of the dataset, and acquisition details of the 1,000 original 3D T1w 3T ADNI scans.

Age (mean (SD))		71.9 (7.0)		
Sex (n, % female)		n = 4644 (46.4%)		
Vendor (%)	GE Medical Systems	20.7%	Discovery MR750	TE = 3.0, TI = 400.0, TR = 7.3
			Signa HDxt	TE = 2.8, TI = 400.0, TR = 7.0
	Philips Medical Systems	19.7%	Achieva	TE = 3.1, TI = 0.0, TR = 6.8
			Gemini	TE = 3.1, TI = 0.0, TR = 6.8
			Ingenia	TE = 3.2, TI = 0.0, TR = 6.8
			Ingenuity	TE = 3.1, TI = 0.0, TR = 6.8
			Intera	TE = 3.2, TI = 0.0, TR = 6.8
	Siemens Medical Solutions	59.6%	Prisma	TE = 3.0, TI = 900.0, TR = 2300.0
			Skyra	TE = 3.0, TI = 900.0, TR = 2300.0
			TrioTim	TE = 3.0, TI = 900.0, TR = 2300.0
			Verio	TE = 3.0, TI = 900.0, TR = 2300.0

Values are reported as mean (standard deviation) for age, as percentages for categorical variables (sex and scanner manufacturer), and in milliseconds for acquisition parameters.

ity levels, indicating that simulated severity corresponded well to the PondrAI QC rating scale (i.e., higher artifact severity was associated with poorer image quality). For *motion*, ratings reported a logarithmic trend, with intermediate motion severities already scoring high on the PondrAI QC scale. *Noise* showed only a slight upward trend, with most ratings remaining between 2 and 4.

While inter-rater reliability was strong when assessed across all artifact types ($\alpha = 0.82$, 95% CI: 0.81–0.83; mean $\rho = 0.87$, 95% CI: 0.84–0.90), it varied considerably between individual artifacts. *Ghosting* showed the highest agreement ($\alpha = 0.79$, 95% CI: 0.77–0.82; mean $\rho = 0.84$, 95% CI: 0.76–0.92). *Blurring* and *motion* had poor agreement based on Krippendorff's α (0.63, 95% CI: 0.59–0.66 and 0.54, 95% CI: 0.49–0.58 respectively), but still relatively high consistency in ranking (mean $\rho = 0.77$, 95% CI: 0.65–0.88; mean $\rho = 0.73$, 95% CI: 0.57–0.88 respectively). For both *noise* and *non-degraded images*, reliability was lower, with Krippendorff's α values of 0.41 (95% CI: 0.35–0.46) and 0.31 (95% CI: 0.13–0.45) and lower mean Spearman's ρ values (0.46, CI: 0.26–0.70 and 0.32, CI: 0.55–0.87, respectively), indicating less agreement in both scores and image ranking. These differences are further illustrated in the pairwise Spearman's ρ heatmaps (Supplementary Figure 1).

3.3. RB, ML, AND DL MODEL COMPARISONS

Detailed RB-, ML-, DL-based classification results for blurring, ghosting, noise, and motion and severity levels 1–10 are shown in Table 2. In terms of computational time, once IQMs had been extracted or the preprocessing had been completed and algorithms were trained, RB, ML, and DL algorithms required approximately 19 ± 6 ms, 3 ± 4 ms, and 2 ± 1 ms per scan, respectively.

For *blurring*, DL-based classification demonstrated strong discrimination across all severities (YI = 0.83–0.97, Figure 3). RB classification showed a strong upward trend,

especially at severity 3, with YI rising from 0.30 to 0.75 and continuing to improve to very strong performance (YI = 0.81) at the highest severity. ML-based performance was also initially weak with YI = 0 at severity 1, but improved rapidly, reaching near-perfect levels (YI ≥ 0.99) by severity 3. For *ghosting*, DL models maintained strong discrimination (YI 0.85–0.92) across all severities, while both RB and classical ML models showed weak discrimination at early severities (YI < 0.15), improving to strong (YI = 0.78) and very strong (YI = 0.92) discrimination, respectively, at maximum severity. For *motion*, DL based classification achieved consistently very strong YI (0.91–0.96), while RB models showed moderate discrimination at low severity (0.47), improving to 0.80 at severity 10, and classical ML based models exhibited a similar trend but with higher maximum performance (YI = 0.97). In the case of *noise*, DL classification maintained strong performance with YI between 0.88 and 0.97, outperforming both RB and ML models at lower severities; classical ML reached perfect discrimination (YI = 1.00) at higher severities, whereas RB models improved gradually from 0.00 to 0.90.

Overall, these findings indicate that DL-based classification is consistently robust across all artifact types and severities, while RB and ML models are more reliably performant only at higher levels of artifact severity and show limited discrimination accuracy at lower severities across all artifact types. Other evaluation metrics, displaying the same trends as the YI, are shown in Supplementary Figures 2–6.

3.4. IQM IMPORTANCE FOR RB- AND ML-BASED CLASSIFICATION

IQMs exhibited distinct patterns across artifact types and severity levels (Supplementary Figure 7, Supplementary Table 2). For *blurring*, there were moderately increased values of EFC ($\rho = 0.68$) and CNR ($\rho = 0.56$), and a weak

Table 2. Performance metrics of rule-based, machine learning-based, and deep learning-based models for the different artifacts and severities.

Artifact	Severity	Balanced Accuracy			Accuracy			F1 Score			Precision			Recall			Youden Index		
		RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL
Blurring	1	0.51	0.50	0.92	0.51	0.50	0.92	0.62	0.51	0.92	0.19	0.49	0.86	0.82	0.52	0.98	0.01	0.00	0.83
Blurring	2	0.65	0.61	0.92	0.65	0.61	0.92	0.70	0.56	0.92	0.51	0.71	0.88	0.79	0.50	0.96	0.30	0.22	0.84
Blurring	3	0.88	0.99	0.98	0.88	0.99	0.98	0.87	0.99	0.98	0.94	0.99	0.97	0.82	1.00	1.00	0.75	0.99	0.97
Blurring	4	0.93	1.00	0.99	0.93	1.00	0.99	0.93	1.00	0.99	1.00	1.00	0.98	0.87	1.00	1.00	0.86	1.00	0.98
Blurring	5	0.90	1.00	0.99	0.90	1.00	0.99	0.89	1.00	0.99	1.00	1.00	0.97	0.80	1.00	1.00	0.80	1.00	0.97
Blurring	6	0.90	1.00	0.98	0.90	1.00	0.98	0.89	1.00	0.98	1.00	1.00	0.97	0.81	1.00	1.00	0.80	1.00	0.97
Blurring	7	0.90	1.00	0.98	0.90	1.00	0.98	0.89	1.00	0.98	1.00	1.00	0.97	0.80	1.00	1.00	0.80	1.00	0.97
Blurring	8	0.89	1.00	0.98	0.89	1.00	0.98	0.87	1.00	0.98	1.00	1.00	0.95	0.77	1.00	1.00	0.77	1.00	0.95
Blurring	9	0.88	1.00	0.98	0.88	1.00	0.98	0.87	1.00	0.98	1.00	1.00	0.96	0.77	1.00	1.00	0.77	1.00	0.95
Blurring	10	0.90	1.00	0.99	0.90	1.00	0.99	0.89	1.00	0.99	1.00	1.00	0.97	0.81	1.00	1.00	0.81	1.00	0.97
Ghosting	1	0.53	0.52	0.92	0.53	0.52	0.92	0.61	0.54	0.93	0.32	0.48	0.85	0.74	0.56	1.00	0.07	0.04	0.85
Ghosting	2	0.57	0.57	0.95	0.57	0.57	0.95	0.64	0.58	0.95	0.38	0.54	0.92	0.76	0.60	0.98	0.14	0.14	0.90
Ghosting	3	0.65	0.65	0.95	0.65	0.65	0.95	0.70	0.66	0.95	0.51	0.62	0.93	0.80	0.68	0.97	0.30	0.30	0.90
Ghosting	4	0.73	0.71	0.95	0.73	0.71	0.95	0.75	0.72	0.95	0.65	0.69	0.91	0.81	0.74	0.99	0.46	0.43	0.90
Ghosting	5	0.80	0.80	0.94	0.80	0.80	0.94	0.81	0.81	0.94	0.78	0.79	0.88	0.83	0.82	1.00	0.61	0.61	0.88
Ghosting	6	0.82	0.87	0.94	0.82	0.87	0.94	0.81	0.87	0.95	0.83	0.87	0.89	0.80	0.87	1.00	0.63	0.73	0.89
Ghosting	7	0.85	0.90	0.95	0.85	0.90	0.95	0.84	0.90	0.95	0.87	0.91	0.91	0.82	0.90	0.98	0.70	0.81	0.89
Ghosting	8	0.89	0.91	0.95	0.89	0.91	0.95	0.88	0.92	0.95	0.93	0.91	0.92	0.85	0.92	0.98	0.78	0.83	0.90
Ghosting	9	0.88	0.94	0.96	0.88	0.94	0.96	0.87	0.94	0.96	0.95	0.91	0.92	0.81	0.96	1.00	0.75	0.87	0.92
Ghosting	10	0.89	0.96	0.96	0.89	0.96	0.96	0.88	0.96	0.96	0.99	0.95	0.93	0.79	0.97	0.99	0.78	0.92	0.92
Motion	1	0.74	0.67	0.96	0.74	0.67	0.96	0.76	0.66	0.97	0.62	0.71	0.93	0.85	0.63	1.00	0.47	0.34	0.93
Motion	2	0.77	0.67	0.96	0.77	0.67	0.96	0.78	0.68	0.96	0.70	0.67	0.91	0.84	0.68	1.00	0.54	0.35	0.91
Motion	3	0.90	0.94	0.98	0.90	0.94	0.98	0.89	0.94	0.98	0.98	0.94	0.96	0.83	0.95	1.00	0.80	0.89	0.95
Motion	4	0.89	0.90	0.96	0.89	0.90	0.96	0.88	0.90	0.96	0.94	0.87	0.93	0.84	0.92	1.00	0.78	0.80	0.92
Motion	5	0.90	0.93	0.96	0.90	0.93	0.96	0.89	0.94	0.96	0.98	0.91	0.93	0.82	0.96	0.99	0.80	0.87	0.91
Motion	6	0.91	0.97	0.95	0.91	0.97	0.95	0.90	0.97	0.95	1.00	0.96	0.92	0.82	0.97	0.99	0.82	0.93	0.91
Motion	7	0.92	0.95	0.96	0.92	0.95	0.96	0.91	0.95	0.96	0.99	0.96	0.94	0.85	0.94	0.99	0.84	0.89	0.93
Motion	8	0.90	0.97	0.96	0.90	0.97	0.96	0.89	0.97	0.96	1.00	0.97	0.94	0.81	0.96	0.98	0.81	0.94	0.92
Motion	9	0.90	0.96	0.97	0.90	0.96	0.97	0.88	0.96	0.97	1.00	0.96	0.95	0.79	0.96	0.99	0.79	0.92	0.94
Motion	10	0.90	0.98	0.98	0.90	0.98	0.98	0.89	0.98	0.98	1.00	0.98	0.96	0.80	0.99	1.00	0.80	0.97	0.96
Noise	1	0.50	0.53	0.94	0.50	0.53	0.94	0.63	0.55	0.94	0.16	0.51	0.90	0.84	0.56	0.97	0.00	0.07	0.88

Artifact	Severity	Balanced Accuracy			Accuracy			F1 Score			Precision			Recall			Youden Index		
		RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL	RB	ML	DL
Noise	2	0.55	0.60	0.97	0.55	0.60	0.97	0.64	0.60	0.97	0.27	0.60	0.97	0.82	0.60	0.97	0.09	0.20	0.94
Noise	3	0.62	0.72	0.97	0.62	0.72	0.97	0.69	0.71	0.98	0.41	0.72	0.95	0.83	0.71	1.00	0.24	0.43	0.95
Noise	4	0.70	0.89	0.97	0.70	0.89	0.97	0.73	0.89	0.97	0.60	0.89	0.95	0.80	0.89	0.99	0.40	0.78	0.94
Noise	5	0.78	0.95	0.97	0.78	0.95	0.97	0.79	0.95	0.97	0.74	0.94	0.94	0.83	0.96	1.00	0.57	0.90	0.94
Noise	6	0.85	0.98	0.98	0.85	0.98	0.98	0.85	0.98	0.98	0.84	0.97	0.95	0.85	0.98	1.00	0.69	0.95	0.95
Noise	7	0.89	0.99	0.97	0.89	0.99	0.97	0.89	0.99	0.97	0.93	0.98	0.94	0.85	0.99	1.00	0.78	0.97	0.94
Noise	8	0.91	1.00	0.97	0.91	1.00	0.97	0.91	1.00	0.97	0.96	0.99	0.94	0.86	1.00	1.00	0.82	0.99	0.94
Noise	9	0.94	1.00	0.97	0.94	1.00	0.97	0.94	1.00	0.98	0.99	1.00	0.95	0.89	1.00	1.00	0.88	1.00	0.95
Noise	10	0.95	1.00	0.98	0.95	1.00	0.98	0.95	1.00	0.98	1.00	1.00	0.97	0.90	1.00	1.00	0.90	1.00	0.97

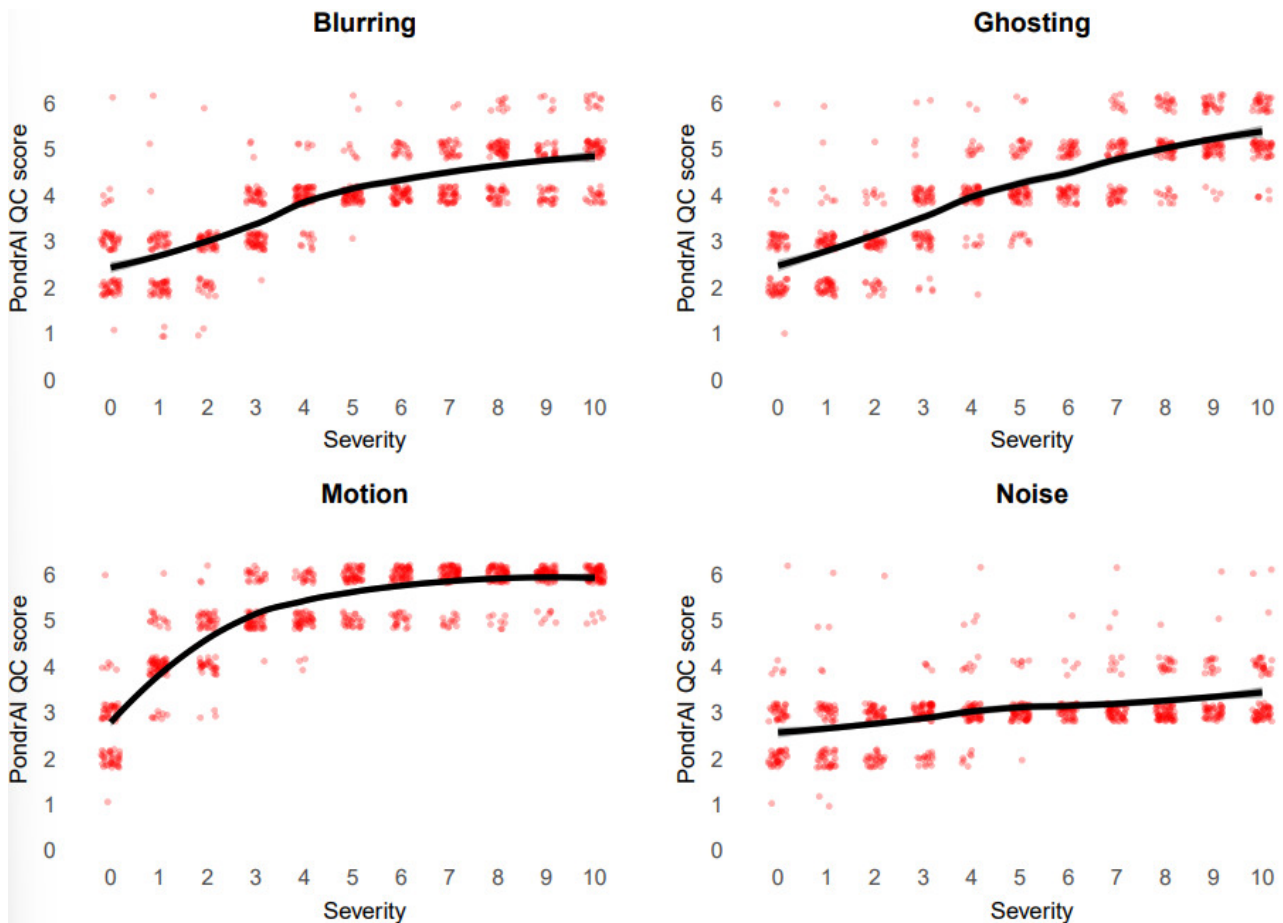


Figure 2. Visual QC ratings based on the PonderAI QC scale (Costantino and Devenyi, 2023) as a function of artifact severity (0–10) for four artifact types: blurring, ghosting, motion, and noise. Red dots show individual ratings; black lines show smoothed trend lines with 95% confidence intervals, computed using locally estimated scatterplot smoothing. Blurring and ghosting display an approximately linear increase in QC ratings with severity, motion shows a ceiling effect, and noise shows minimal change.

positive association for TPM ($\rho = 0.33$), alongside a weak negative association for INU ($\rho = -0.24$), indicating greater image entropy, enhanced contrast, and tissue probability map distortion at higher severities. *Ghosting* led to weak increases in INU ($\rho = 0.35$) and weak declines in TPM ($\rho = -0.31$), while EFC and CNR remained relatively stable ($\rho = -0.15$ and $\rho = 0.05$, respectively), suggesting spatial misalignment without major changes in contrast or entropy. *Motion* was associated with weakly increasing INU ($\rho = 0.34$) and moderately decreasing TPM ($\rho = -0.42$), reflecting spatial distortion and misclassification of tissue classes, while EFC and CNR remained stable. In contrast, *noise* resulted in declines in EFC ($\rho = -0.71$) and moderate increases in INU ($\rho = 0.50$), while showing negligible-to-weak declines in both CNR ($\rho = -0.21$) and TPM ($\rho = -0.14$). These trends highlight the complementary sensitivity of EFC, CNR, INU, and TPM to different artifact types and intensities (Supplementary Figure 7).

The pattern of IQM selection across artifact types and severity levels for the RB and ML-based classification is displayed in Figure 4. For *blurring*, both ML and RB models initially selected a broad mix of normative and entropy IQMs, alongside some image contrast IQMs. From severity level

3 onward, RB and ML models consistently prioritized normative and entropy IQMs, particularly those reflecting spatial properties and deviations in cerebrospinal fluid (CSF), while rule-based models reduced their selection to a smaller set of normative features, mainly related to CSF and spatial structure. For *ghosting*, RB models relied mainly on normative IQMs related to the gray matter and white matter, and a few image contrast IQMs. ML models relied at lower severities mainly on normative IQMs related to the CSF, while at higher severities it shifted to normative IQMs related to the white matter, and some image contrast IQMs. For *motion*, both RB and ML models initially used normative IQMs. However, RB models' most informative normative IQMs were related to gray matter, white matter, and CSF, whereas ML models used mainly background related normative IQMs. At higher severities, the most informative IQMs for RB models were the normative IQMs. ML models relied on the GM and background related normative IQMs and SNR-related image contrast IQMs. For *noise*, ML models initially drew from all four categories but shifted toward normative IQMs, particularly those reflecting CSF, gray matter, and entropy measures (full width at half maximum related). In contrast, RB models narrowed their focus

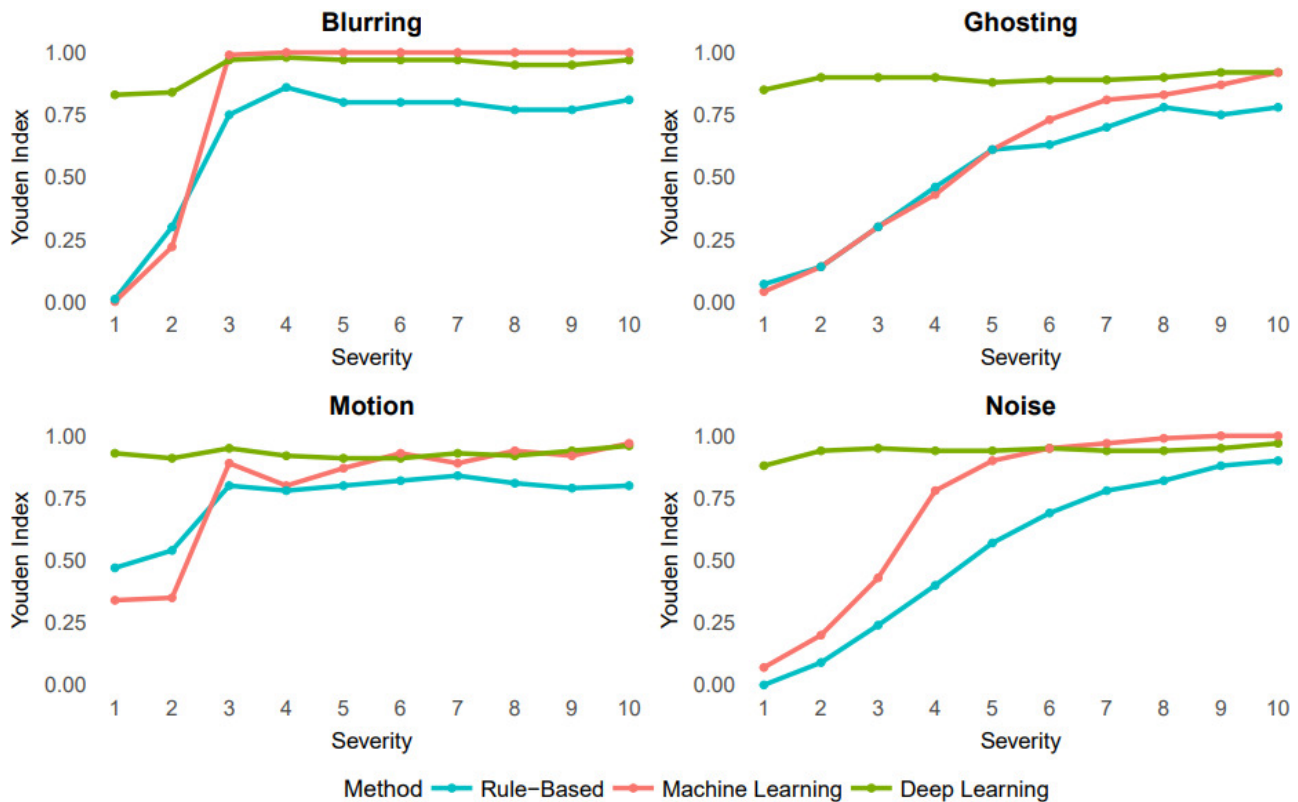


Figure 3. Youden Index scores across artifact severity levels (1–10) for each artifact type (blur, ghosting, motion, and noise), comparing performance of rule-based (blue), machine learning-based (red), and deep learning-based (green) classification methods. Each panel presents results for a single artifact type. Lines represent the Youden Index (sensitivity + specificity – 1) at each severity level.

more abruptly, retaining mainly normative (GM related) and image contrast (SNR related) IQMs.

4. DISCUSSION

This benchmark study yielded three main findings. First, while visual QC ratings correlated with simulated artifact severity, the non-linearity of these relations varied between the artifacts. Second, while there was little performance difference for detecting high severity simulated artifacts, DL outperformed RB and ML for detecting subtle simulated artifacts. RB and ML performed similarly, with performance varying between artifact types. Third, RB models relied mainly on a consistent core of normative features with limited use of contrast measures, whereas ML models adapted their feature use depending on artifact type and severity, drawing more flexibly from multiple IQM categories.

Our findings highlight that human visual QC is not equally effective across artifact types, which may have important implications for its use as a benchmark. The high inter-reader agreement for ghosting suggests that artifacts of this type can be reliably perceived. For blurring and motion, inter-rater agreement was lower but individual raters' assessments remained relatively monotonic with respect to the simulated level of the artifact. Noise, in particular, was rated differently by different readers, perhaps reflecting its weaker perceptual salience, or limited noise vari-

ability in the higher severities of the PondrAIQC training examples. Another potential explanation for the relative insensitivity of human raters to noise is that they can consistently interpret detail in an image obscured with noise.³⁸ These findings are in line with previous studies. When comparing human visual QC with machine-based, non-artifact specific, QC scores, previous studies also found poor correspondence, particularly for subtle or borderline real-world artifacts.^{39,40} Another study with synthetic artifacts found stronger correlations of radiologist diagnostic quality scores with perceptually aligned image quality metrics, such as DISTs and HaarPSI, than with other IQMs.⁴¹ Together, this highlights the limitations of visual QC as a gold standard for subtle (synthetic) artifacts, whether for detecting general image quality or for detecting specific artifacts.

The fact that DL mostly outperformed RB and ML at low artifact severities aligns with previous studies, which suggest that DL architectures can learn more complex hierarchical image representations and more subtle textural and structural deviations than traditional image features.^{13–15, 42} Within our benchmark, this advantage seems especially pronounced for simulated artifacts with distributed, non-local changes in image intensity and spatial coherence, such as the simulated ghosting and motion.^{10,43} Nevertheless, RB and ML approaches remained highly effective in detecting simulated blurring and noise artifacts, especially at higher severity levels, where degradation can be more easily captured by single or composite IQMs.^{12,44} In line

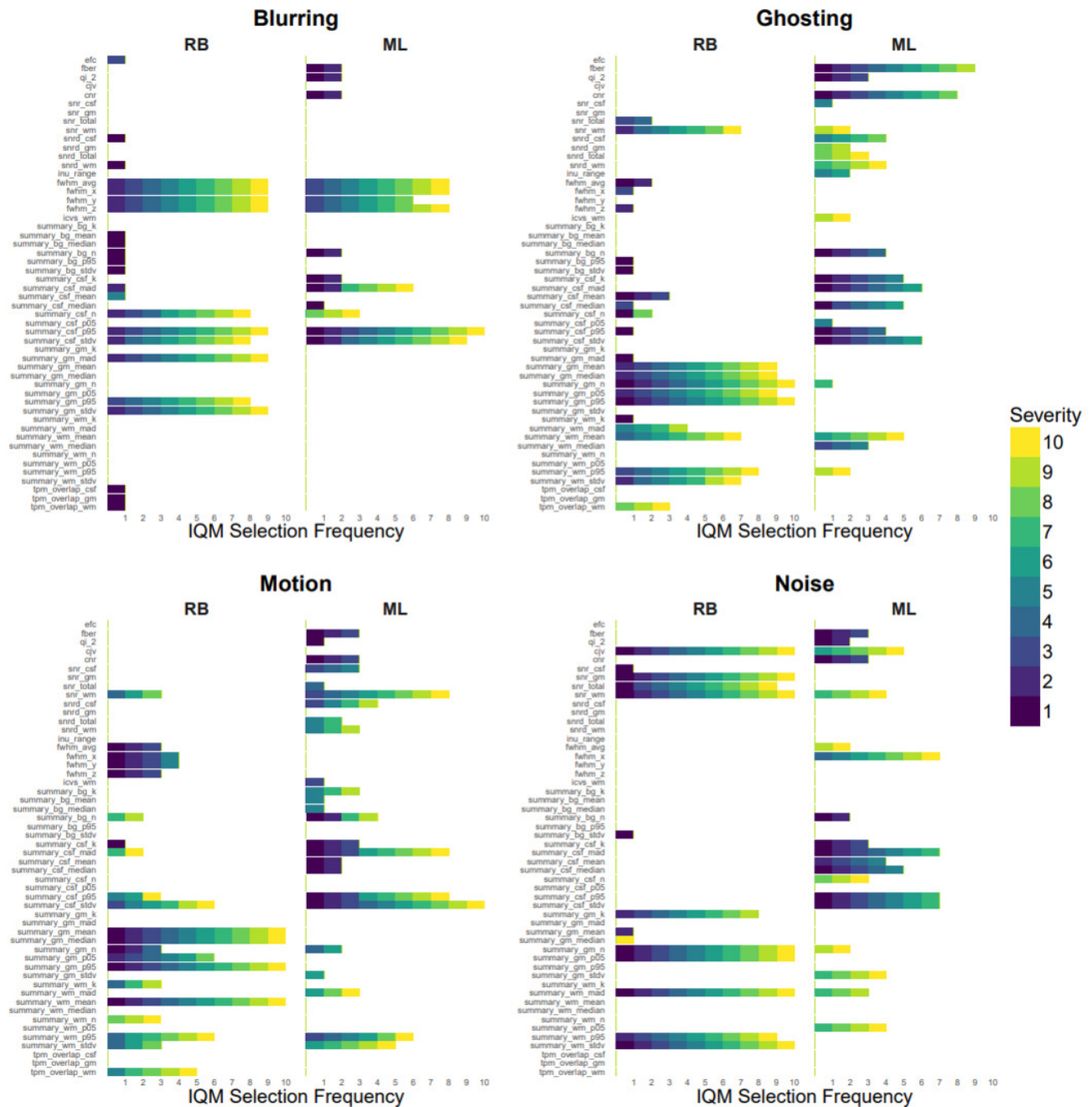


Figure 4. Frequency of IQM selection across artifact types, severity levels and classification models. Each panel shows the number of models (x-axis) that selected each IQM (y-axis) for a given artifact type (blur, ghosting, motion, noise), separately for rule-based (RB) and machine learning-based (ML) classification. Color indicates artifact severity (1–10). As the deep learning-based classification models did not have input features, they are not shown in this figure.

with this, it has been shown that, in head-to-head comparison, classical ML models trained on robust IQMs can match DL models for motion artifacts.⁴⁵ Together with these studies, our findings suggest that simpler ML models might offer an alternative to DL, which can be beneficial as they can be more efficient and easier to implement, especially when the goal is to merely detect gross artifacts.

The importance of IQMs for simulated artifact detection by RB and ML varied with the simulated artifact type and severity. At low artifact severities, both RB and ML relied on a broad spectrum of normative and image contrast IQMs to detect subtle or diffuse degradations. With increasing

severity, the IQM selection consistently shifted toward normative features, particularly those quantifying deviations from the expected anatomical patterns and spatial properties. Similarly, previous studies demonstrated that global entropy and contrast IQMs are sensitive to early or mild artifacts,^{9,44,46} while normative, region-based IQMs offer superior sensitivity and specificity for more severe or localized degradations.^{10,39} The increasing reliance on normative IQMs when using simulated artifacts might be relevant in large neuroimaging studies, where heterogeneity in acquisition and (physio-)pathology can obscure subtle artifacts unless deviations from normative patterns are explic-

itly modeled.^{9,10,40} However, it remains to be established the same feature-selection patterns occur for real-world artifacts.

A key strength of our study is the use of synthetic artifacts for benchmarking, which allows for precise control over artifact type and severity, and the ability to generate large, balanced datasets for model training and evaluation.^{23,24} However, these simulations should be regarded as controlled approximations rather than realistic reproductions of real-world MRI artifacts, as the simulations do not capture the heterogeneous and subtle degradations that arise from the complex interactions between head motion, hardware imperfections, and acquisition parameters.^{42,47} In particular, the artifacts were simulated on magnitude-only reconstructed images rather than using complex-valued raw k-space data. Consequently, the simulations do not reproduce the full interaction between acquisition-time artifacts, coil sensitivities, channel combination, parallel-imaging reconstruction, and vendor-specific preprocessing. Our simulated motion uses simplified perturbations of image-derived k-space and does not model motion confined to subsets of k-space, so it captures only a subset of motion-artifact phenotypes. Similarly, the simulated ghosting is generated as shifted copies of the signal along a single axis and does not include reconstruction-related failures or all aliasing-like patterns. For blurring, a Gaussian image-space blur cannot represent all real-world sources of blurring. The simulated noise is additive, zero-mean Gaussian noise without spatial correlation and therefore does not reproduce the spatially varying, non-Gaussian noise that can result from multichannel reconstruction and vendor preprocessing. We did not validate the results based on simulated artifacts against independently labelled real-world motion, noise, ghosting, or blurring artifacts. Therefore, the present findings establish performance only under the specific simulated artifacts used here and cannot be assumed to generalize to real-world artifacts.

The ADNI preprocessing and reconstruction pipeline includes gradient-nonlinearity distortion correction, B1 and N3 intensity normalization, phantom-based spatial scaling, and vendor-applied gradient-linearity and B1-corrections (<https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/mri-pre-processing/>). These steps improve geometric and intensity homogeneity but place our simulations in a preprocessed image-space domain rather than in the native acquisition domain. Therefore, image-space simulations such as blurring, ghosting, motion, and noise may differ from corresponding artifacts as they arise in real-world raw k-space, and the effects of the ADNI reconstruction and preprocessing pipeline may have removed, attenuated, or transformed artifact characteristics before simulation. This aligns with broader concerns in MRI benchmarking where the use of processed or simplified data can produce inflated performance estimates and reduced generalizability to real raw acquisitions.⁴⁸ As a result, the algorithms we trained on our synthetic artifacts may have partially exploited regularities introduced by the simulation pipeline itself, rather than fully capturing the heterogeneous and subtle degradations typically observed in real-

world artifacts, and therefore we cannot determine whether the relative performance of RB, ML, and DL would be preserved when detecting real-world artifacts in images produced by different acquisition and reconstruction pipelines.

Moreover, real MRI scans often exhibit multiple, overlapping artifact types that can impact image quality simultaneously, perhaps often at too-low severities to be visible to the human eye. We chose only to compare the detectability of individual simulated artifacts to maximize interpretability. With the present data, we cannot determine how the interaction of multiple artifacts would interfere with the detection of a single specific artifact, nor whether such interference matters if the primary goal is to assess overall image quality rather than identify a specific artifact type. The models may behave differently when artifact types interact. For these reasons, our conclusions are restricted to the relative performance of QC strategies in detecting individual synthetic artifacts on ADNI T1w MPRAGE scans, and therefore the performance metrics reported should be interpreted as characterizing relative behavior within this controlled benchmark and should not be reused quantitatively for clinical QC deployment without further validation.

Recent work employed transfer learning to retrain DL models on small sets of real, artifact-labeled images after pretraining on synthetic artifacts,²⁴ showing that this combination of simulated and real-world artifacts leads to higher artifact detection performance and generalizability. Although this combination of simulated and real-world artifacts led to higher performance and generalizability, it does not allow disentangling the presence of individual artifacts. Automatic disentangling between artifact types may have value to differentiate scanner and sequence stability from patient compliance. Together, these findings suggest that future research should employ advanced explainable DL algorithms, trained with a combination of individual synthetic artifacts and real-world artifacts.

Another factor limiting the generalizability of our study is the reliance on a single dataset for both training and validation. Although ADNI is a multisite study, its standardized protocols and participant demographics may not fully represent the diversity encountered in clinical practice or other research cohorts.^{9,11,16} This homogeneity could inflate model performance and limit its generalizability. Previous studies have highlighted the challenges of cross-site generalization for automated QC, emphasizing the need for diverse, multi-cohort benchmarks and harmonization strategies.^{10,42,49} Future studies should prioritize external validation using datasets with real-world artifacts and broader demographic variability.

Furthermore, we prioritized comparability over performance, by employing ML and DL models that were often used and available open-source, with their standard parameter settings from previous studies. While these models ranked among the best-performing approaches for T1w QC,^{10,14,39,40,42} they may be outperformed by more recent and advanced methods, such as a ProtoPNet⁵⁰ or by more extensive hyperparameter optimization or ensemble methods. Additionally, we did not investigate why DL models

were more successful at detecting subtle artifacts, as exploring model interpretability (e.g., through saliency mapping, SHAP values, or feature attribution methods),^{13,51} was beyond the scope of this study. Systematically benchmarking advanced DL models and integrating explainable AI in future studies could establish the true upper bounds of automated QC performance in neuroimaging.

Finally, the relevance of a QC score and the ability to define a QC threshold may vary between use cases. For example, a radiologist may still find a subtly-degraded image diagnostically usable while it already has lost its potential to show subtle findings in group studies. Training DL models may require different image quality than image-derived phenotypes. Future studies may investigate if QC scores predict the robustness of image-derived phenotypes, affect statistical power, and how automatic QC can be used in clinical practice or trial workflows.

In conclusion, our findings demonstrate that, within this controlled synthetic artifact benchmark, DL offers superior performance for the automated detection of individual synthetic artifacts in T1w brain MRI, particularly for subtle artifacts. RB and ML methods may remain valuable for interpreting quality degradation of artifacts with well-defined quantifiable features. For RB and ML methods, the varying importance of IQMs across simulated artifact types and severities highlight the need for flexible, multi-metric QC frameworks. Future studies should address the challenges of real-world generalizability, model complexity, and interpretability between synthetic and real-world artifacts to fully realize the potential of automated QC in neuroimaging.

CODE AVAILABILITY

The code for artifact simulation and the training of the rule-based algorithm is available at the Corresponding author's GitHub:

<https://github.com/janinehendriks-hub/AutoQC-compare-MRI>

ACKNOWLEDGEMENTS

This project was funded by Health~Holland, Top Sector Life Sciences & Health (TKI-PPP Amsterdam UMC 2011227).

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Submitted: March 31, 2026 CDT. Revised: June 03, 2026 CDT; August 24, 2026 CDT. Accepted: October 07, 2026 CDT. Published: October 27, 2026 CDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

REFERENCES

1. Alexander-Bloch A, Clasen L, Stockman M, et al. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Hum Brain Mapp*. 2016;37(7):2385-2397. doi:[10.1002/hbm.23180](https://doi.org/10.1002/hbm.23180)
2. Perrone D, Aelterman J, Pizurica A, Jeurissen B, Philips W, Leemans A. The effect of Gibbs ringing artifacts on measures derived from diffusion MRI. *Neuroimage*. 2015;120:441-455. doi:[10.1016/j.neuroimage.2015.06.068](https://doi.org/10.1016/j.neuroimage.2015.06.068)
3. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*. 2012;59(3):2142-2154. doi:[10.1016/j.neuroimage.2011.10.018](https://doi.org/10.1016/j.neuroimage.2011.10.018)
4. Reuter M, Tisdall MD, Qureshi A, Buckner RL, van der Kouwe AJW, Fischl B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage*. 2015;107:107-115. doi:[10.1016/j.neuroimage.2014.12.006](https://doi.org/10.1016/j.neuroimage.2014.12.006)
5. Birn RM. Quality control procedures and metrics for resting-state functional MRI. *Frontiers in neuroimaging*. 2023;2:1072927. doi:[10.3389/fnimg.2023.1072927](https://doi.org/10.3389/fnimg.2023.1072927). PMID:37554646
6. Taylor PA, Glen DR, Reynolds RC, Basavaraj A, Moraczewski D, Etzel JA. Demonstrating quality control (QC) procedures in fMRI. Published online 2023:1205928. doi:[10.3389/978-2-8325-2704-7](https://doi.org/10.3389/978-2-8325-2704-7)
7. Scheltens P, Launer LJ, Barkhof F, Weinstein HC, Van Gool WA. Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: interobserver reliability. *Journal of neurology*. 1995;242(9):557-560. doi:[10.1007/BF00868807](https://doi.org/10.1007/BF00868807)
8. Williams B, Hedger N, McNabb CB, Rossetti GM, Christakou A. Inter-rater reliability of functional MRI data quality control assessments: A standardised protocol and practical guide using pyfMRIqc. *Frontiers in Neuroscience*. 2023;17:1070413. doi:[10.3389/fnins.2023.1070413](https://doi.org/10.3389/fnins.2023.1070413). PMID:36816136
9. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*. 2018;166:400-424. doi:[10.1016/j.neuroimage.2017.10.034](https://doi.org/10.1016/j.neuroimage.2017.10.034)
10. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One*. 2017;12(9):e0184661. doi:[10.1371/journal.pone.0184661](https://doi.org/10.1371/journal.pone.0184661)
11. Hendriks J, Mutsaerts HJ, Joules R, et al. A systematic review of (semi-)automatic quality control of T1-weighted MRI scans. *Neuroradiology*. 2024;66(1):31-42. doi:[10.1007/s00234-023-03256-0](https://doi.org/10.1007/s00234-023-03256-0)
12. Kim H, Irimia A, Hobel SM, et al. The LONI QC System: A Semi-Automated, Web-Based and Freely-Available Environment for the Comprehensive Quality Control of Neuroimaging Data. *Front Neuroinform*. 2019;13:60. doi:[10.3389/fninf.2019.00060](https://doi.org/10.3389/fninf.2019.00060)
13. Bottani S, Burgos N, Maire A, et al. Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. *Med Image Anal*. 2022;75:102219. doi:[10.1016/j.media.2021.102219](https://doi.org/10.1016/j.media.2021.102219)
14. Fantini I, Yasuda C, Bento M, Rittner L, Cendes F, Lotufo R. Automatic MR image quality evaluation using a Deep CNN: A reference-free method to rate motion artifacts in neuroimaging. *Comput Med Imaging Graph*. 2021;90:101897. doi:[10.1016/j.compmedimag.2021.101897](https://doi.org/10.1016/j.compmedimag.2021.101897)
15. Keshavan A, Yeatman JD, Rokem A. Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Front Neuroinform*. 2019;13:29. doi:[10.3389/fninf.2019.00029](https://doi.org/10.3389/fninf.2019.00029)
16. Kruggel F, Turner J, Muftuler LT, Alzheimer's Disease Neuroimaging I. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *Neuroimage*. 2010;49(3):2123-2133. doi:[10.1016/j.neuroimage.2009.11.006](https://doi.org/10.1016/j.neuroimage.2009.11.006)
17. Krupa K, Bekiesinska-Figatowska M. Artifacts in magnetic resonance imaging. *Pol J Radiol*. 2015;80:93-106. doi:[10.12659/PJR.892628](https://doi.org/10.12659/PJR.892628)
18. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*. 2010;74(3):201-209. doi:[10.1212/WNL.0b013e3181cb3e25](https://doi.org/10.1212/WNL.0b013e3181cb3e25)
19. Jack CR Jr, Arani A, Borowski BJ, et al. Overview of ADNI MRI. *Alzheimers Dement*. 2024;20(10):7350-7360. doi:[10.1002/alz.14166](https://doi.org/10.1002/alz.14166)

20. Manning EN, Leung KK, Nicholas JM, et al. A comparison of accelerated and non-accelerated MRI scans for brain volume and boundary shift integral measures of volume change: evidence from the ADNI dataset. *Neuroinformatics*. 2017;15(2):215-226. doi:[10.1007/s12021-017-9326-0](https://doi.org/10.1007/s12021-017-9326-0). PMID:28316055
21. Fernandez-Lozano S, Dadar M, Morrison C, et al. QRATER: a collaborative and centralized imaging quality control web-based application. *bioRxiv*. Published online 2022:2022.12.20.521204. doi:[10.1101/2022.12.20.521204](https://doi.org/10.1101/2022.12.20.521204)
22. Costantino M, Devenyi G. PONDRAI QC. 2024. <https://github.com/pondrai/PondrAIQC>
23. Perez-Garcia F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed*. 2021;208:106236. doi:[10.1016/j.cmpb.2021.106236](https://doi.org/10.1016/j.cmpb.2021.106236)
24. Loizillon S, Bottani S, Maire A, et al. Automatic motion artefact detection in brain T1-weighted magnetic resonance images from a clinical data warehouse using synthetic data. *Med Image Anal*. 2024;93:103073. doi:[10.1016/j.media.2023.103073](https://doi.org/10.1016/j.media.2023.103073)
25. Richard Shaw CS, Ourselin S, Cardoso MJ. MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty. *Proceedings of Machine Learning Research*. Published online 2019:427-436.
26. Santiago AFGVSF. *Statistical Analysis of Noise in MRI*. Springer International Publishing; 2016. doi:[10.1007/978-3-319-39934-8](https://doi.org/10.1007/978-3-319-39934-8)
27. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. Sage publications; 2018. doi:[10.4135/9781071878781](https://doi.org/10.4135/9781071878781)
28. Nogueira S, Sechidis K, Brown G. On the use of Spearman's rho to measure the stability of feature rankings. Published online 2017:381-391. doi:[10.1007/978-3-319-58838-4_42](https://doi.org/10.1007/978-3-319-58838-4_42)
29. de Carvalho M, Marques FJ. Jackknife euclidean likelihood-based inference for Spearman's rho. *North American Actuarial Journal*. 2012;16(4):487-492. doi:[10.1080/10920277.2012.10597644](https://doi.org/10.1080/10920277.2012.10597644)
30. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 1996;29(3):162-173. doi:[10.1006/cbmr.1996.0014](https://doi.org/10.1006/cbmr.1996.0014)
31. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M. SynthStrip: skull-stripping for any brain image. *Neuroimage*. 2022;260:119474. doi:[10.1016/j.neuroimage.2022.119474](https://doi.org/10.1016/j.neuroimage.2022.119474)
32. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62(2):782-790. doi:[10.1016/j.neuroimage.2011.09.015](https://doi.org/10.1016/j.neuroimage.2011.09.015)
33. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008;12(1):26-41. doi:[10.1016/j.media.2007.06.004](https://doi.org/10.1016/j.media.2007.06.004)
34. Shehzad Z, Giavasis S, Li Q, et al. The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data. *Frontiers in Neuroscience Conference Abstract: Neuroinformatics 2015*. Published online 2015.
35. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:[10.1007/s10994-006-6226-1](https://doi.org/10.1007/s10994-006-6226-1)
36. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320. doi:[10.1109/TMI.2010.2046908](https://doi.org/10.1109/TMI.2010.2046908)
37. Akoglu H. User's guide to correlation coefficients. *Turkish journal of emergency medicine*. 2018;18(3):91-93. doi:[10.1016/j.tjem.2018.08.001](https://doi.org/10.1016/j.tjem.2018.08.001). PMID:30191186
38. Simonotto E, Riani M, Seife C, Roberts M, Twitty J, Moss F. Visual perception of stochastic resonance. *Physical review letters*. 1997;78(6):1186. doi:[10.1103/PhysRevLett.78.1186](https://doi.org/10.1103/PhysRevLett.78.1186)
39. Pizarro RA, Cheng X, Barnett A, et al. Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm. *Front Neuroinform*. 2016;10:52. doi:[10.3389/fninf.2016.00052](https://doi.org/10.3389/fninf.2016.00052)
40. Rosen AFG, Roalf DR, Ruparel K, et al. Quantitative assessment of structural image quality. *Neuroimage*. 2018;169:407-418. doi:[10.1016/j.neuroimage.2017.12.059](https://doi.org/10.1016/j.neuroimage.2017.12.059)
41. Kastruyulin S, Zakirov J, Pezzotti N, Dylov DV. Image quality assessment for magnetic resonance imaging. *IEEE Access*. 2023;11:14154-14168. doi:[10.1109/ACCESS.2023.3243466](https://doi.org/10.1109/ACCESS.2023.3243466)

42. Bhalerao G, Gillisa G, Dembele M, et al. Automated quality control of T1-weighted brain MRI scans for clinical research datasets: methods comparison and design of a quality prediction classifier. *Imaging Neuroscience*. 2025;3. doi:[10.1162/IMAG.a.4](https://doi.org/10.1162/IMAG.a.4)
43. Shaw R, Sudre C, Ourselin S, Cardoso MJ. MRI k-space motion artefact augmentation: model robustness and task-specific uncertainty. Published online 2019.
44. Mortamet B, Bernstein MA, Jack CR Jr, et al. Automatic quality assessment in structural brain magnetic resonance imaging. *Magn Reson Med*. 2009;62(2):365-372. doi:[10.1002/mrm.21992](https://doi.org/10.1002/mrm.21992)
45. Vakli P, Weiss B, Szalma J, et al. Automatic brain MRI motion artifact detection based on end-to-end deep learning is similarly effective as traditional machine learning trained on image quality metrics. *Med Image Anal*. 2023;88:102850. doi:[10.1016/j.media.2023.102850](https://doi.org/10.1016/j.media.2023.102850)
46. Woodard JP, Carley-Spencer MP. No-reference image quality metrics for structural MRI. *Neuroinformatics*. 2006;4(3):243-262. doi:[10.1385/NI.4.3.243](https://doi.org/10.1385/NI.4.3.243)
47. Smith TBN, Krishna S. MRI artifacts and correction strategies. *Imaging in Medicine*. 2010;2(4):445-457. doi:[10.2217/iim.10.33](https://doi.org/10.2217/iim.10.33)
48. Shimron E, Tamir JI, Wang K, Lustig M. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*. 2022;119(13):e2117203119. doi:[10.1073/pnas.2117203119](https://doi.org/10.1073/pnas.2117203119). PMID:35312366
49. White T, Jansen PR, Muetzel RL, et al. Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction. *Hum Brain Mapp*. 2018;39(3):1218-1231. doi:[10.1002/hbm.23911](https://doi.org/10.1002/hbm.23911)
50. Garcia M, Dosenbach N, Kelly C. BrainQCNet: a Deep Learning attention-based model for the automated detection of artifacts in brain structural MRI scans. *Imaging Neuroscience*. 2024;2:1-16. doi:[10.1162/imag_a_00300](https://doi.org/10.1162/imag_a_00300). PMID:40800355
51. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. 2019;29(2):102-127. doi:[10.1016/j.zemedi.2018.11.002](https://doi.org/10.1016/j.zemedi.2018.11.002)

SUPPLEMENTARY MATERIALS

Supplementary Material

Download: <https://apertureneuro.org/article/172318-automatic-detection-of-simulated-artifacts-on-t1w-magnetic-resonance-images-comparing-performance-of-different-qc-strategies/attachment/370695.docx>
