Received: 5 June 2018 | Revised and accepted: 12 October 2018

DOI: 10.1002/hbm.24449

WILEY

RESEARCH ARTICLE

Supervised machine learning quality control for magnetic resonance artifacts in neonatal data sets

Yang Ding[1,2] | Sabrina Suffren[1,2] | Pierre Bellec[2,3,4] | Gregory A. Lodygensky[1,2,5]

[1]Department of Pediatrics, Sainte-Justine University Hospital and University of Montreal, Montreal, Quebec, Canada

[2]Canadian Neonatal Brain Platform, Montreal, Quebec, Canada

[3]Centre de Recherche, Institut Universitaire de Gériatrie de Montréal Montreal, Montreal, Quebec, Canada

[4]Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada

[5]Department of Pharmacology and Physiology, Sainte-Justine University Hospital and University of Montreal, Montreal, Quebec, Canada

Correspondence
Gregory Lodygensky, Department of Pediatrics, Neonatology, University of Montreal and Ste-Justine University Hospital, 3175 Côte Ste-Catherine, Montréal, Québec, H3T 1C5, Canada.
Email: ga.lodygensky@umontreal.ca

Funding information
Fondation Brain Canada, Grant/Award Number: 7108124

Abstract

Quality control (QC) of brain magnetic resonance images (MRI) is an important process requiring a significant amount of manual inspection. Major artifacts, such as severe subject motion, are easy to identify to naïve observers but lack automated identification tools. Clinical trials involving motion-prone neonates typically pool data to obtain sufficient power, and automated quality control protocols are especially important to safeguard data quality. Current study tested an open source method to detect major artifacts among 2D neonatal MRI via supervised machine learning. A total of 1,020 two-dimensional transverse T2-weighted MRI images of preterm newborns were examined and classified as either QC Pass or QC Fail. Then 70 features across focus, texture, noise, and natural scene statistics categories were extracted from each image. Several different classifiers were trained and their performance was compared with subjective rating as the gold standard. We repeated the rating process again to examine the stability of the rating and classification. When tested via 10-fold cross validation, the random undersampling and adaboost ensemble (RUSBoost) method achieved the best overall performance for QC Fail images with 85% positive predictive value along with 75% sensitivity. Similar classification performance was observed in the analyses of the repeated subjective rating. Current results served as a proof of concept for predicting images that fail quality control using no-reference objective image features. We also highlighted the importance of evaluating results beyond mere accuracy as a performance measure for machine learning in imbalanced group settings due to larger proportion of QC Pass quality images.

KEYWORDS

brain imaging, Canadian Neonatal Brain Platform, motion detection, neonatal, open source, quality control, T2w

1 | INTRODUCTION

Modern medical imaging analysis pipeline such as brain segmentation, cortical morphometry analysis requires high quality data as even subtle motion has shown to bias measurements (Alexander-Bloch et al., 2016; Reuter et al., 2015). The implementation of quality control (QC) process ensures the accuracy of brain imaging measurements. This is particularly important for multicenter clinical trials involving neonates and infants, where pooling data from multiple local cohorts is often required to obtain sufficiently powered studies. In addition, neonates and pediatric images typically face additional challenges such as motion and noncooperative patients during scanning sessions (Raschle et al., 2012). To address these challenges, the Canadian Neonatal Brain Platform brought together a national, multidisciplinary team of researchers and clinicians to identify causes of brain dysmaturation and develop strategies to minimize brain injury occurring during the neonatal period (https://www.cnbp.ca). However, merging data from different institutions mandates standardization including QC process and site-specific acquisition protocol. These needs highlight the importance of scalability and automation.

Machine learning has been successfully implemented across multiple fields, especially in the context of image object recognition where it has even achieved classification results exceeding human accuracy (Ren, He, Girshick, & Sun, 2015). The goal of the current work was to

devise and empirically test a modular and scalable workflow pipeline that adapts to user input to perform semiautomated quality assessments of MRI image data using a supervised machine learning approach to help facilitating larger scale rating and QC of neuroimaging data sets in the future. The most similar categories of image analyses based QC approach is the Preprocessed Connectome Project Quality Assessment Protocol (PCP-QAP, see http://preprocessed-connectomes-project.org/quality-assessment-protocol/) and its derived Stanford's MRIQC initiative (https://github.com/poldracklab/mriqc and http://mriqc.readthedocs.io) (Esteban et al., 2017). Both pipelines are heavily optimized toward high quality 3D anatomical and 4D functional neuroimaging research data sets. Both tools measure features such as noise, spatial prior distribution of information, statistical properties of tissue distributions, and the sharpness/blurriness of the images. Still, there is a demand for a similar QC approach on routinely acquired lower resolution 2D clinical data sets with little or no anatomical prior and brain segmentation to assist with the isolation and identification of artifacts.

To this end, we tailored our approach to 2D images acquired in clinical settings instead. In addition, we evaluated existing image-processing and analyses algorithms to postulate a series of objective features from independent single 2D image. Currently four major categories of image-derived features were extracted per image: (1) focus and blurriness measurements; (2) signal, noise and signal-to-noise ratio measurements; (3) texture analyses mostly through statistical summaries of the gray level codependence matrix (GLCM); and (4) natural scene statistics measurements (see Supporting Information Method section for details). These four categories of features were proposed because each potentially contributed to measure different aspects of unique properties of the images. Focus features were chosen mostly to help detecting motion induced onionskin-like artifacts (as typically seen in Figure 1 and Supporting Information Figure S1). This resulted in less sharp contrasting edges and defocusing of the image; the signal and noise ratio detection features were mainly based on the mean signal intensity of the central region of the image with no other prior anatomical assumptions. MRI signals are acquired in radio frequency domains hence aberrant signals can result in recurring wave like patterns when images are reconstructed in spatial domain resulting in features that may not be adequately described by aforementioned focus or noise features. These repetitive patterns within the image may be detected using texture features via adjacency measures such as those using GLCM. Lastly, organic shapes have typically smoother shapes in comparison to sharp edges of man-made objects or distortion artifacts. Natural scene statistics were shown to emphasize organic and smooth edges/shapes over preferences to more commonly seen artificial structures and shapes (Moorthy & Bovik, 2011; Sheikh, Bovik, & De Veciana, 2005).

2 | METHODS

2.1 | Participants

Following ethical approval, we selected a cohort of preterm neonates with bronchopulmonary dysplasia imaged for clinical purpose at term equivalent age scanned between March 31, 2009 and March 4, 2012. All these neonates were hospitalized at the Centre Hospitalier Universitaire Sainte-Justine (CHU Sainte-Justine) for premature birth below 29 weeks of gestational age and underwent an axial T2w clinical MRI on the same scanner with identical parameters.

2.2 | MRI acquisition parameters

All clinical images were acquired on a Siemens (Munich, Germany) Avanto 1.5 T scanner (Device Serial 25603) with console version Syngo MR B15/B17 at CHU Sainte-Justine. These T2w Axial turbo spin echo images were acquired with repetition time (TR) of 6,480 ms, echo time (TE) of 106 ms, across 20 slices each with 4 mm thickness in ascending acquisition orders. Field of view was adjusted per patient varies between 119 mm by 159 mm to 157 mm by 180 mm. All relevant subject data were exported after subject anonymization from Synapse PACS system (Fujifilm Medical System USA, Stamford, Connecticut). Across the 51 subjects, 1,020 images were successfully imported.

2.3 | Quality assessment

2.3.1 | Subjective rating

Subjective manual image quality assessment ratings were obtained first. All T2w images were sorted by anatomical positions from inferior to superior (for ease of comparison of similar section across subjects) and then visually checked and rated individually for quality control by a coauthor (SS) without awareness of the subsequent in silico results. The image quality of each individual 2D image was classified by the rater as either QC Pass or QC Fail, where QC Fail typically indicated significant artifacts such as severe subject motion or acquisition issues (see the complete list of examples in supporting information figure S1). In our subsequent comparisons between objective machine classifications and subjective ratings, subjective ratings were considered as the gold standard. Two separate subjective ratings were carried out by the same coauthor (SS) and Cohen's kappa value was calculated to assess intra-rater reliability.

2.3.2 | Objective features

Four main categories of objective features were obtained per 2D image: Focus, signal to noise, texture, and natural scene statistics. Even within each category, different features characterize the image quality from different perspectives: for instance, both spatial frequency approach (Eskicioglu & Fisher, 1995) and wavelet domain approach (Xie, Rong, & Sun, 2006) can be used to characterize image focus.

- Focus features provided quantitative measurements to the blurriness within the image set to detect motion artifacts or aliasing during typical acquisition processes (Pertuz, Puig, & Garcia, 2013). A total of 25 focus measurement features were computed based on an adapted version of Pertuz's code.
- Noise measures were included to compensate the focus features because high contrast noise (e.g., salt and pepper noise) would result in elevated measure of focus, despite being another indicator of poor image quality. In addition, overly noisy images could be indicative of poor MRI data. Fernandez et al. reviewed and implemented nine algorithms to estimate Rician noise typically found in 2D magnetic

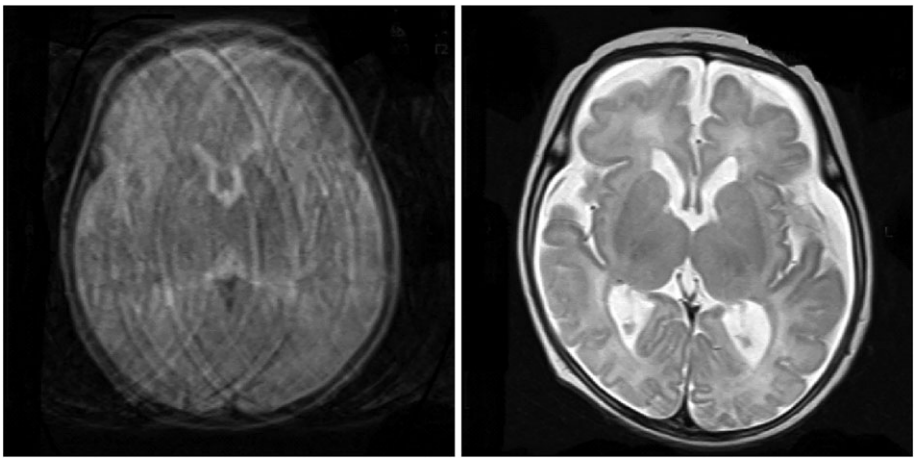

FIGURE 1 Example of images that fail (left) and pass (right) subjective quality control

resonance imaging (Aja-Fernandez, Tristan-Vega, & Alberola-Lopez, 2009). An adapted version of their code was included. We also included of six signal related measurements. For signal measurement, since without prior information, it is impossible to identify the precise region of interest to be considered "signal" with respect to noise, the average signal intensities from the central region of the 2D image (50×50 pixels or 100×100 pixels) were used as a crude proxy measure of the real signal for the anatomical region. This was done purposely to help generalize this process beyond brain quality control and avoid using anatomical priori information such as brain segmentation with the implicit assumption that the regions of interest were more likely to be in the middle of the image. The signal measurements were then compared with the noise measurement to derive the signal to noise ratio.

- Texture analyses in the form of GLCM, a measure of the statistical relationship between two adjacent pixels, were also implemented by adapting Dr. Avinash Uppuluri's approach (Uppuluri, 2008). These measures complement existing focus features (by highlighting elevated degree of correlation with neighboring pixels) as well as noise measure (MRI Rician noise is more random and less likely to possess texture and patterns).
- Lastly, seven natural scene statics (NSS) features were included. Natural scene statistics are measurements of the physical regularities within images that depict and stem from natural environment. Natural scene evolutionarily influences the development of early stage of neuronal processing specialization notably in the visual cortex (Olshausen & Field, 1996) and help facilitating the innate ability to differentiate between natural versus manmade objects in the environments. Almost all NSS features were developed from the University of Texas at Austin, Laboratory for Image & Video Engineering labs (http://live.ece.utexas.edu/research/quality), namely: Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) from (Mittal, Moorthy, & Bovik, 2012), Blind Image Quality Index (BIQI) from (Moorthy & Bovik, 2010). NSS for compressed image (JP2KNR) by Sheikh, Bovik, and Cormack (2005), Spatial Spectral Entropy based Quality features (SSEQ) by (Liu, Liu, Huang, & Bovik, 2014) and Naturalness Image Quality Evaluator (NIQE) by (Mittal et al., 2012) have been implemented. MatLab's implementation of BRISQUE and BIQI were also additionally included. We implicitly assume biomedical images such as brain images should conform as natural scenes with smoother shapes and curvatures more so than shapes observed in distortion, motion or inhomogeneity acquisition artifacts.

Further details about individual features, including relevant references, are included in the Supporting Information Material.

The overall features extraction pipeline was built to be modular and expandable to accommodate additional features in the future. In the current analyses, 70 features were extracted per input 2D image.

2.4 | Analyses of quality assessment results

Once the subjective ratings and objective image features were derived, two steps of analyses were implemented (see Figure 2). The first analysis step was to inspect the various quantitative values of the objective features using traditional statistics and check if (and how) they differed across the subjectively classified rating groups. This was important as the objective image quality ratings and subjective ratings were independently derived. While the objective features were chosen purposefully to ideally represent the essential image features underlying the subjective ratings, it was not guaranteed, and the relation between these two sources of data needed to be statistically validated (see Figure 2, Top).

The second analysis step was to apply supervised machine learning to classify images into groups based on the input from the objective image features and subjective labels. This step was important because even if statistical differences across the subjective rating groups were established, they cannot guarantee it was possible to precisely classify the group membership at individual 2D image level. Ultimately a good quality control process must be able to efficiently and effectively discriminate the images with good accuracy, sensitivity and specificity into QC Pass versus QC Fail labels.

2.4.1 | Step 1: Statistical analyses

For all statistical analyses (see Figure 2, Top), a family-wise alpha threshold of 0.05 was set a priori as significant. All statistically analyses were performed in SPSS 23.0 (SPSS, Chicago, Illinois).

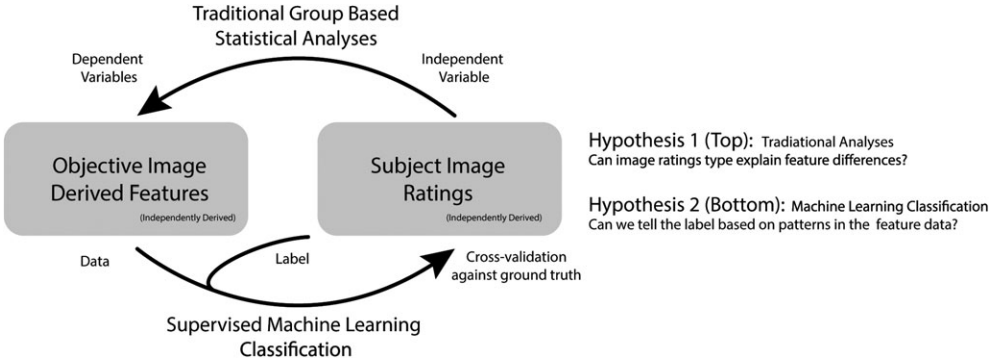

FIGURE 2 Graphical illustration of the analyses strategy and respective hypotheses

First, an omnibus one-way MANOVA test was conducted with subjective manual classified ratings (i.e., QC Pass or QC Fail) as independent variable and the various objective features as dependent variables. This omnibus MANOVA test could reveal (with family-wise error corrected) if overall, subjective ratings were associated with the differences observed in objective image features.

If and only when such significant differences were detected, further post-hoc one-way omnibus MANOVA tests of specific category of objective image features (i.e., focus or SNR or texture or NSS) were evaluated to determine if the subjectively rating groups differed in terms of the major category of features. The alpha for this post-hoc omnibus was family-wise corrected via Bonferroni correction and set to be at 0.05/4 categories = 0.0125.

Further exploration of individual feature within each category of objective image features were conducted via post-hoc univariate *t*-test with a Bonferroni corrected alpha level and were presented in the Supporting Information Material.

2.4.2 | Step 2: Supervised machine learning

To further explore our subjective quality assessment and to take advantage of these objective features (see Figure 2, Bottom), several different supervised machine learning models were trained and tested with the end goal of producing a machine model that can accurately classify image quality based on the training input. This was carried out using MATLAB version 2017b with Statistics and Machine Learning toolbox (MathWorks, Natick, MA). The analysis was conducted with 10-fold cross-validation and we tested different kinds of classifiers including linear and quadratic discriminant analyses which were common approaches for initial analyses, naïve Bayes Classifier and decision tree (Breiman, Friedman, Stone, & Olshen, 1984) which were insensitive to irrelevant features, and Random UnderSampling and adaBoost (RUSBoost) ensemble method (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2008) which specifically alleviated the extreme ratings in-balance observed in our dataset (Table 1). A mock illustration (not using our data) of how these classifiers differ in their higher-level approach to classifying 2D data is given in the Supporting Information Figure S2. Because of the rating imbalance in the data, three key measures were focused on: (1) positive predictive value (true QC fail images/predicted QC fail images) which represented the proportion of classifier predicted poor images that were actually images with poor qualities; (2) sensitivity (true QC fail images/all QC failed images) which represented the ability to detect all known poor quality images; and (3) f1 score, the harmonic mean of the above two measures. By choosing these key features to gauge the performance of the classifiers instead of accuracy, we avoided classification performance bias from imbalanced data groups. Accuracy in our example would not be representative of the performance across both groups since QC pass images represent about 95% of all our data. We also calculated the Cohen's kappa between the subjective rating and the best performing machine learning model.

2.5 | Computation performance benchmark

The predictive analyses involved two major processes: the objective image features extraction and the model training process, both were benchmarked. To compute all 70 image features on a six core Xeon CPU E5-2620 2.00 Ghz HP z620 workstation with 32 GB RAM running Windows 10, it took about 2.5 hr across all 1,024 images. Once all features were extracted, to generate and train all supervised classifier models and to summarize and report their model performance data took 10 min.

Once the model is trained and saved to disk, we also benchmarked the time requirement for QC prediction on a new image. Prediction involves two similar phases: (1) passing the input 2D image through the entire metrics extraction pipeline, which took 1.5 min mostly due to Matlab parallel processing pool setup overhead, and (2) apply the best trained classifier to predict QC status of the images based on its extracted features which required 0.12 s.

2.6 | Source code

All the image-based features extraction, classification analyses, including results and MatLab source code were uploaded to http://github.com/CNBP/DICOMetrics. An anonymized version of the extracted

TABLE 1 Similarity matrix between first and second classification of the same dataset

		First rating		
		QC fail	QC pass	Total
Second rating	QC fail	44	7	51
	QC pass	15	954	969
Total		59	961	1,020

QC = quality control.

TABLE 2 MANOVA Wilk's lambda test measuring the amount of variance observed in the four image feature categories not accounted for by the variance in the subjective manual ratings

Feature type	Wilk's lambda	*F*	Hypothesis DF	Error DF	*p*	Partial Eta squared
Focus and Blurr	0.11	342.5	24	995	<.001	0.89
Signal and noise	0.87	10.3	14	1,005	<.001	0.13
Texture	0.54	44.4	19	1,000	<.001	0.46
Natural scene statistics	0.56	115.5	7	1,012	<.001	0.44
Combination of all feature types	0.08	201.2	56	963	<.001	0.92

DF = degree of freedoms.

70 features along with their QC labels (without any identifiable information) was also available in the Results folder. Future production branches of our software can utilize smaller sets of metrics once the final prediction model is fully validated across wider range of datasets and populations.

3 | RESULTS

3.1 | Subjective manual quality assessment

The results of the repeated subjective rating of the images into QC Pass or QC Fail are shown in Table 1. Overall, our subjective ratings demonstrated considerable consistency across the two blind subjective ratings of the same set of images. Cohen's κ were calculated to determine if there were substantial agreement between the two quality assessments of the 1,020 two-dimensional MRI images. There was a substantial agreement as defined in Landis and Koch (1977), between the two ratings of the same image sets, $\kappa = 0.79$ (95% CI: 0.70–0.87), $p < .0005$. The results based on second repeated rating included in the Supporting Information Result.

3.2 | Objective image features quality assessment

Seventy features were derived per image across four major categories: 25 features to measure image focus, 15 features to measure signal/noises, 23 features to measure image texture, and seven features to measure natural scene statistics.

Detailed distribution of features was posted on DICOMetric Github pages "Results" section. Overall, we observed strong asymmetry and bimodal distributions across majority of the focus and SNR features, whereas the natural scene statistics and texture features showed a unimodal and normal distributions across them.

3.3 | Statistical analyses of quality assessment results

3.3.1 | All objective image feature categories

The omnibus one-way MANOVA test across all 70 features between subjective ratings (QC Pass vs. QC Fail) has shown that there was overall a statistically significant difference among all image features between QC Pass and QC Fail rated images ($F[58, 961] = 198.6$, $p < .0005$; Wilk's $\Lambda = 0.077$, partial $\eta_2 = 0.923$). This warranted further post hoc analyses to examine, among all four feature categories, which specific category of features significant differs between the QC Pass and QC Fail rated images.

3.3.2 | Individual objective image feature category

When the post hoc MANOVA analysis was conducted on each individual feature category, the results showed statistically significant differences between the two subjective ratings across every single individual feature category (Table 2): focus features, signal/noise features, texture features, and natural scene statistics features, where all *p* values were far below the 0.0125 Bonferroni corrected alpha threshold.

Notably, focus features category overall had the lowest value for Wilk's Λ (Table 2), where Wilk's Λ value was used to indicate the proportion of variance in dependent variables not explained by the independent variables. Here, low Wilk's Λ value suggested relatively smaller amount of variance in the focus image features (dependent variables) were unexplained by variance in QC Pass versus QC Fail subject rating assignment (independent variable). Other feature categories such as noise, texture and natural scene statistics all had Wilk's Λ values much higher than 0.5 suggesting majorities of variance in these image features were not well explained by the subjective rating assignment, despite statistical significance.

These results established that images with different subject ratings were statistically significantly different in terms of their objectively derived image features (Figure 2, Top, Hypothesis 1). However, such group level statistical significance in features cannot guarantee effective case by case classification performance at individual image level (Figure 2, Bottom, Hypothesis 2).

We also included detailed individual objective image feature comparison across QC Pass and QC Fail groups in the Supporting Information Tables S1 and S2.

3.3.3 | Supervised machine learning and quality assessment results

The results of the effective predictive power of objectively derived image features to infer the appropriate subjective rating at individual image level are summarized in Table 3. In short, we observed overall better performance from decision tree based approaches (decision tree and RUSBoost ensemble) than discriminant and naïve Bayes classifiers. Overall, the prediction accuracies of all classifiers were elevated because of excessively large number of QC Pass images (95 + %, See Table 1) and hence the classification algorithms performance were biased toward QC Pass images in most scenarios at the expense of worse performance toward the QC Fail

TABLE 3 Classification performance of various classifiers tested

Performance (10-fold cross validation)	Precision (%)	Recall (%)	F1 score	Negative predictive value (%)	Specificity (%)	Overall prediction accuracy (%)
Linear discriminant analyses	60.7	62.7	61.67	97.7	97.5	95.5
Quadratic discriminant analyses	46.2	81.4	58.90	98.8	94.2	93.4
Naïve Bay classifier	78.4	49.2	60.42	96.9	99.2	96.3
Decision tree	66.1	69.5	67.77	98.1	97.8	96.2
RUSBoost (min false negative)	65.8	84.7	74.07	99.0	97.3	96.6
RUSBoost (min false positive)	84.6	74.6	79.28	98.5	99.2	97.7

Min = minimizing; RUSBoost, Random UnderSampling and adaBoost.

images. Given such premises, we chose best classifier based on its performance of positive predictive value and sensitivity with respect to QC Fail images. The best classifier was the RUSBoost decision tree with a false positive emphasis that on average after tenfold cross-validation resulted in over 84% positive predictive value with about 75% sensitivity and F1 score of 0.79 for QC Fail images. The Cohen's kappa between final model prediction versus the first subjective manual rating was 0.78 (95% CI: 0.70–0.87) $p < .0005$, on par with kappa of intra-rater reliability.

4 | DISCUSSION

Current work proposed a novel set of quality control processes for 2D image by a features extraction pipeline producing 70 features across four major categories: image focus, signal/noise, adjacent pixel relation/texture, and natural scene statistics. Our statistical analyses (Figure 2, Top) first showed that the two subjective rating groups (QC Pass and QC Fail) differed across all four categories of image features. Therefore, these features were relevant, informative and potentially useful features in the machine learning analyses (Figure 2, Bottom).

The MANOVA statistical analysis section and Table 3 highlighted several important results: (1) Taken as a collective measure, the objective features overall significantly differed between the subjective QC rating groups, (2) when further examined, every individual category of image features (i.e., focus, SNR, texture or natural scene statistics) were significantly different between the subjective rating groups, (3) when evaluating the proportion of the variance explained using MANOVA models, the subjective rating differences were better at explaining the variance observed in focus features (as revealed by a small Wilk's $\Lambda = 0.11$), as compare to the other three categories of features (where Wilk's $\Lambda > 0.5$). These findings supported the application of a supervised machine learning approach (specifically, a classification scenario) by using image features to infer the individual subjective ratings.

Several common models of supervised machine learning classifiers were tested, namely: linear discriminant analysis, quadratic discriminant analysis, naïve Bayes classifier, decision tree, and variants of Random UnderSampling and adaBoost (RUSBoost) decision tree ensemble models. After utilizing 10-fold cross validation to mitigate model overfitting, the end results showed that RUSBoost decision tree ensemble model performed the best in identifying QC Fail images with a sensitivity of 75% and positive predictive value of 80%, replicated across second subjective rating (Supporting Information Table S3 and Supporting Information Figure S3). The reason RUSBoost tree performed best was most likely accounted by its special design to deal with classifications of imbalanced groups (Seiffert et al., 2008). Other classification approaches in general typically performed well in either sensitivity (quadratic discriminant analyses, RUSBoost with false negative focus) or positive predictive value (linear discriminant analyses, decision tree) but seldom both. Overall, this was an encouraging result given that across the 1,020 images, only 59 images were bad images, accounting approximately 5% of the dataset, yet the supervised machine learning algorithms were able to achieve a reasonable sensitivity and positive predictive value after 10-fold cross-validating the results. In fact, the machine learning predicted labels achieving similar agreement ($\kappa = 0.78$) as the intra-rater reliability on the same dataset ($\kappa = 0.79$). On the other hand, for the QC Pass cases, our classifier performed very well as expected since there were substantially more QC Pass images due to class imbalance in the source data.

Quality control has always been a quintessential process before any MRI data analyses and yet to this date no objective measures or formalized process have been established due to diverse nature of the MRI image acquisition protocols. To ensure a consistent and generic quality assurance process, it is imperative to design an objective image quality metric evaluation process, that is, robust, time efficient, and highly quantitative yet flexible enough to adapt to different needs. We ultimately decided to choose a 2D metric extraction approach for the following reasons:

1. *Generalizability*: Our current extraction and validation pipeline is not protocol specific. It does not require any acquisition protocol to be adapted as long as the subjective rating data set is congruent with testing data set.
2. *Independence of anatomical prior*: Our approach does not require brain segmentation, in fact, none of our analyses even assumed there was a brain in the DICOM file. As long as the training dataset is appropriate, the classifiers will mimic how subjective rating classifies the image qualities regardless of the content of the images or even criteria of the classification process.
3. *Interoperability*: We have designed our quality control pipeline with clinical images in mind where typically there is a much shorter acquisition time, larger anisotropic voxel size due to

thicker slices and often with additional interslice gaps. By working at individual 2D DICOM image level, we do not require high-resolution 3D acquisitions, yet our approach is still forward-compatible with high quality scans allowing future possible application to both typical clinical and research protocols. Secondly, this quality control process was completely decoupled from any subsequent image analyses/examinations. It can provide the researchers/clinicians with more granular objective and quantifiable information about the number of images (and proportion of images in 3D volume data) that had image quality issues.

4. *Extensibility*: while current version output 70 features, there is no limit on the number of features that can be extracted in the future and incorporated. The only limitation lies with the subsequent negligible increase in training speed of the classifiers.

4.1 | Limitations/outlook

The improvement of classification performance will likely be achieved by continuously adding more target data sets (i.e., QC Fail). With more extensive input of larger labeled poor-quality data, we can further extend the generalizable performance of these classifiers.

Current level of classification performance is most likely higher than its performance on completely new unseen data set. Our current results are inherently relying on implicit assumptions these few QC-Fail images are representative of all QC-Fail images in the real-world scenario. Although we maximized compatibility with clinical usage by drawing these data directly from MRI scanners using typical clinical acquisitions sequences, these QC fail images are very unlikely to represent the full spectrum of all the poor-quality images, emphasizing the importance and benefit of having sufficiently largely labeled data for all the necessary classes. It is our hope that with more extensive input of labeled poor-quality data such as those from open science quality control dataset and collective expert input (Keshavan, Yeatman, & Rokem, 2018), we can further extend the performance and generalization of these classifiers.

5 | CONCLUSION

Our long-term development roadmap for DICOMetrics is to provide a comprehensive quality assessment and quantification toolsets that are nonbiased, objective and useful. To achieve that, we started by working with 2D image features because they are the common basis of both structural (typically 3D high spatial resolution) as well as functional or diffusion EPI images (typically 4D low spatial resolution). Our current results showcased a proof-of-concept open source pipeline to help delineate 2D T2 MRI images with major quality issues. The pipeline operated at individual 2D DICOM image level and leveraged existing no-reference image features to produce 70 image features across four major categories. The results indicated a promising potential for supervised machine learning based identification of poor-quality images with a positive prediction ratio above 85% and sensitivity above 75% with no anatomical prior, segmentation or reference high quality images. Our pipeline will be proudly incorporated as part of the Canadian Neonatal Brain Platform to facilitate the quality control process of all neonatal researchers Canada wide and further improve its performance and generalizability.

ACKNOWLEDGMENT

The authors would like to acknowledge Ms. Genevieve Blain for help with data acquisition and procurement. This work was supported by the Brain Canada Foundation (Canada) that funded the Canadian Neonatal Brain Platform (https://cnbp.ca).

ORCID

Yang Ding https://orcid.org/0000-0003-1004-5555

REFERENCES

Aja-Fernandez, S., Tristan-Vega, A., & Alberola-Lopez, C. (2009). Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models. *Magnetic Resonance Imaging*, *27*(10), 1397–1409. https://doi.org/10.1016/j.mri.2009.05.025

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., & Raznahan, A. (2016). Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human Brain Mapping*, *37*(7), 2385–2397. https://doi.org/10.1002/hbm.23180

Breiman, L., Friedman, J., Stone, C., & Olshen, R. A. (1984). *Classification and regression trees (Wadsworth statistics/probability)*. Boca Raton, FL: Chapman and Hall/CRC.

Eskicioglu, A. M., & Fisher, P. S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications*, *43*(12), 2959–2965. https://doi.org/10.1109/26.477498

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *Plos One*, *12*(9), e0184661.

Keshavan, A., Yeatman, J., & Rokem, A. (2018). Combining citizen science and deep learning to amplify expertise in neuroimaging. *bioRxiv*. https://doi.org/10.1101/363382

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Liu, L., Liu, B., Huang, H., & Bovik, A. C. (2014). No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, *29*(8), 856–863. https://doi.org/10.1016/j.image.2014.06.006

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. https://doi.org/10.1109/Tip.2012.2214050

Moorthy, A. K., & Bovik, A. C. (2010). A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, *17*(5), 513–516. https://doi.org/10.1109/Lsp.2010.2043888

Moorthy, A. K., & Bovik, A. C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, *20*(12), 3350–3364.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network*, *7*(2), 333–339. https://doi.org/10.1088/0954-898X/7/2/014

Pertuz, S., Puig, D., & Garcia, M. A. (2013). Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, *46*(5), 1415–1432. https://doi.org/10.1016/j.patcog.2012.11.011

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., … Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Annals of the New York Academy of Sciences*, *1252*, 43–50. https://doi.org/10.1111/j.1749-6632.2012.06457.x

Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster R-CNN: Towards real-time object detection with region proposal networks*. Paper presented at the Advances in Neural Information Processing Systems.

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). *RUSBoost: Improving classification performance when training data is skewed*. Paper presented at the 2008 19th International Conference on Pattern Recognition, ICPR 2008, Tampa, FL.

Sheikh, H. R., Bovik, A. C., & Cormack, L. (2005). No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, *14*(11), 1918–1927. https://doi.org/10.1109/TIP.2005.854492

Sheikh, H. R., Bovik, A. C., & De Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, *14*(12), 2117–2128.

Uppuluri, A. (2008). GLCM texture features. Retrieved from https://www.mathworks.com/matlabcentral/fileexchange/22187-glcm-texture-features?s_tid=prof_contriblnk

Xie, H., Rong, W. B., & Sun, L. N. (2006). *Wavelet-based focus measure and 3-D surface reconstruction method for microscopy images*. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ding Y, Suffren S, Bellec P, Lodygensky GA. Supervised machine learning quality control for magnetic resonance artifacts in neonatal data sets. *Hum Brain Mapp*. 2019;40:1290–1297. https://doi.org/10.1002/hbm.24449