# A Deep Learning–Based Approach to Reduce Rescan and Recall Rates in Clinical MRI Examinations

A. Sreekumari, D. Shanbhag, D. Yeo, T. Foo, J. Pilitsis, J. Polzin, U. Patil, A. Coblentz, A. Kapadia, J. Khinda, A. Boutet, J. Port, and I. Hancu

## ABSTRACT

**BACKGROUND AND PURPOSE:** MR imaging rescans and recalls can create large hospital revenue loss. The purpose of this study was to develop a fast, automated method for assessing rescan need in motion-corrupted brain series.

**MATERIALS AND METHODS:** A deep learning–based approach was developed, outputting a probability for a series to be clinically useful. Comparison of this per-series probability with a threshold, which can depend on scan indication and reading radiologist, determines whether a series needs to be rescanned. The deep learning classification performance was compared with that of 4 technologists and 5 radiologists in 49 test series with low and moderate motion artifacts. These series were assumed to be scanned for 2 scan indications: screening for multiple sclerosis and stroke.

**RESULTS:** The image-quality rating was found to be scan indication– and reading radiologist–dependent. Of the 49 test datasets, technologists created a mean ratio of rescans/recalls of (4.7 ± 5.1)/(9.5 ± 6.8) for MS and (8.6 ± 7.7)/(1.6 ± 1.9) for stroke. With thresholds adapted for scan indication and reading radiologist, deep learning created a rescan/recall ratio of (7.3 ± 2.2)/(3.2 ± 2.5) for MS, and (3.6 ± 1.5)/(2.8 ± 1.6) for stroke. Due to the large variability in the technologists' assessments, it was only the decrease in the recall rate for MS, for which the deep learning algorithm was trained, that was statistically significant ($P = .03$).

**CONCLUSIONS:** Fast, automated deep learning–based image-quality rating can decrease rescan and recall rates, while rendering them technologist-independent. It was estimated that decreasing rescans and recalls from the technologists' values to the values of deep learning could save hospitals $24,000/scanner/year.

**ABBREVIATIONS:** CB = clinically bad; CG = clinically good; CNN = convolutional neural network; DL = deep learning; D0–D5 = radiologists; IQ = image quality; R0 = radiologist; ROC = receiver operating characteristic; T1–T4 = MR imaging technologists

MR imaging is the preferred approach for diagnosing neurologic disorders due to its versatile contrast. Its high cost, however, limits its use. It was recently discovered that repeat acquisitions can significantly extend MR imaging examination time and increase hospital costs. Up to 20% of MR imaging examinations have at least a repeat series, leading to the loss to a hospital of ~$115,000/scanner/year.[1]

Series are repeated when the scanning technologist decides that image quality (IQ) is inadequate for diagnosis. A related problem, wherein a patient is sent home with the technologist assessing that the IQ is sufficient and then recalled due to a radiologist's inability to diagnose, also exists. Reducing rescans and recalls is important for optimizing the efficiency of the health care system. This problem, however, is not easy to solve, as it is the radiologist who decides if IQ is sufficient, and the technologist who makes the rescan decision. In addition, reports document different radiologists' opinions regarding IQ[2] or diagnosing disease.[3] It is likely that a given IQ level may be sufficient for a given physician and insufficient for another.

A few publications exist, documenting means for automated IQ assessments in MR imaging.[2,4-8] In most reports, the IQ of specific imaging sequences (such DWI,[5,6] or of particular acquisitions scanned for a cohort study[2,7]) is assessed. Time-intensive preprocessing, such as brain tissue classification or registration,[2]

Received August 15, 2018; accepted after revision November 2.

From the GE Global Research Center (A.S., D.S., U.P.), Bangalore, India; GE Global Research Center (D.Y., T.F., I.H.), Niskayuna, New York; Albany Medical College (J.Pilitsis), Albany, New York; GE Healthcare (J.Polzin), Milwaukee, Wisconsin; University Hospital Network (A.C., A.K., J.K., A.B.), Toronto, Ontario, Canada; and Mayo Clinic (J.Port), Rochester, Minnesota.

This work was prepared while Ileana Hancu was employed at GE Global Research Center Niskayuna, NY 12309. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Please address correspondence to Ileana Hancu, PhD, GE Global Research Center, 1 Research Circle, K-1 NMR 138, Niskayuna, NY 12309; e-mail: ihancu1@gmail.com

Indicates article with supplemental on-line photos.

http://dx.doi.org/10.3174/ajnr.A5926

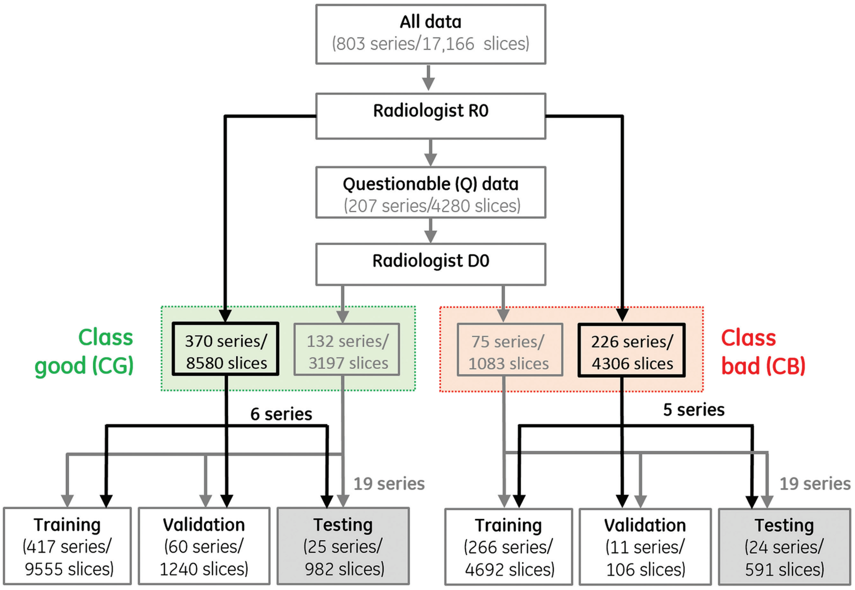

**FIG 1.** Workflow for image rating and usage.

or the extraction of many image features,[4] precludes real-time IQ determination. Faster machine and deep learning (DL)–based methods for artifact detection in MR images have also been reported, generally resulting in per-patch or per-slice classification.[2,4,9-14] While artifact detection on a per-patch or per-slice basis is helpful, it does not inform the technologist about whether a series needs to be rescanned. In many instances, artifacts present in select slices do not require a series rescan.

The fundamental goal of this article was to develop a real-time approach for helping technologists decide whether a series needs to be rescanned. Image quality of individual brain slices of any contrast, pathology, or orientation is first assessed by a DL architecture. Individual slice ratings are subsequently used to compute a per-series score, which is compared against a threshold to decide whether the series requires a rescan. This threshold can be adjusted to accommodate different clinical scan indications and reading radiologists. The performance of this algorithm is validated against assessments from multiple MR imaging technologists and radiologists and for multiple scan indications.

## MATERIALS AND METHODS

### Training, Validation, and Testing Data

This retrospective study was approved by the institutional review board of Albany Medical College. Brain examinations from patients scanned on three HDx 1.5T scanners (GE Healthcare, Milwaukee, Wisconsin) were used for training the DL-based approach. Data were intentionally enriched in questionable/poor IQ series. While series with any type of artifact were initially accepted in the study, it was found that motion was the dominant cause of artifacts (~95%). Due to the lack of data, only motion-corrupted datasets were included in training/validation/testing. Anatomic images of all orientations, all contrast types, and all pathologies (strokes/mass occupying lesions/multiple sclerosis, etc) for cartesian *k*-space sampling schemes were included. Due to the lack of sufficient training sets, DWI datasets were excluded. A good algorithm to identify motion in DWI, based on the phase-striping pattern of moving subjects, has already been published.[5]

Data were initially rated by a single radiologist (R0) with 20 years of experience and partitioned into clinically good (CG), clinically bad (CB), and questionable (Q). The questionable series were then passed to a second radiologist, D0, with 36 years of experience, who, by using the clinical indication of MS, classified them as CG and CB. This 2-tiered rating (exemplified in Fig 1) was undertaken to enable testing in predominantly questionable datasets. Making the right rescan decisions in such cases is truly relevant because very good/bad datasets can be easily classified. Figure 1 also summarizes the data partitioning into training/validation/testing sets; 30%/23%/24%/23% of the (original) datasets used for training belonged to the T1/T2/FLAIR/T2* categories, respectively. To better balance the number of training images in the CB and CG classes, hence classification accuracy, zooms and translations of the initial 4692 images existent in the 266 CB training series were also performed, generating a total of 7783 CB training slices.

On-line Fig 1 presents image examples that were initially rated by R0 as CG/CB (upper/lower rows). The middle rows of this figure represent images that were initially rated as questionable by R0 and ultimately moved by D0 to CG (second row)/CB (third row), respectively.

### Deep Learning–Based Classification

A 2D classification model, implemented in Chollet[15] with a Tensorflow backend was trained to classify individual MR imaging slices into CG or CB. The architecture of the convolutional neural network (CNN), with 7 convolutional layers, 4 max pooling layers, and 3 batch normalization layers (Fig 2), was inspired by ResNet (https://its.unc.edu/resource/resnet/).[16] The number of filters for convolution layers 1 through 7 are 32/64/32/64/32/128/256, while the filter sizes are $(3 \times 3)/(5 \times 5)/(3 \times 3)/(5 \times 5)/(3 \times 3)/(5 \times 5)/(5 \times 5)$, respectively. The activation function, a nonlinear exponential linear unit, helps learn complex patterns from data. To enhance dominant features, 2 merge layers were introduced using the multiplication operation. At the end of all the convolutional layers, a "flatten" layer was used, which converts the feature tensor from convolutional layers to a 1D tensor. A "tanh" activation was followed by a fully connected layer and "softmax" output.[17] The fully connected layer further helps learn the nonlinear combinations of features provided by CNN layers. The softmax function provides probabilities for each class, with the sum of the probabilities equaling 1. Categoric cross-entropy was used as a loss function, and the optimizer was set to "RMSprop".[15] All images used for training and testing were converted to a 128 × 128 size. Pixel values were transformed into *z* score maps, defined as $(Pixel_{\text{value}} - Mean[Series])/Standard\_Deviation(Series)$. Training and testing of the model were performed on a 7910 workstation (Dell, Austin, Texas; 48 CPU cores with an NVIDIA P5000 GPU card). The size of the trained Keras model (h5 format) was ~8 MB.

The DL model outputs the probability for each slice belonging to the CG class. Because rescan decisions are made on a per-series, not per-slice, basis, individual slice ratings were pooled to compute a per-series score, expressed as the geometric mean of the per-slice probabilities ($P\ [Series] = \sqrt{P1 \times P2 \times\ ...\ \times Pn}$, where $P1, P2, ... Pn$ are predictions for slices 1, 2, . . . $n$). Finally, a series is rated as CG if $P\ (Series) \geq t$, and as CB if $P\ (Series) < t$, where $t$ is a threshold that can vary as a function of scan indication and reading radiologist.

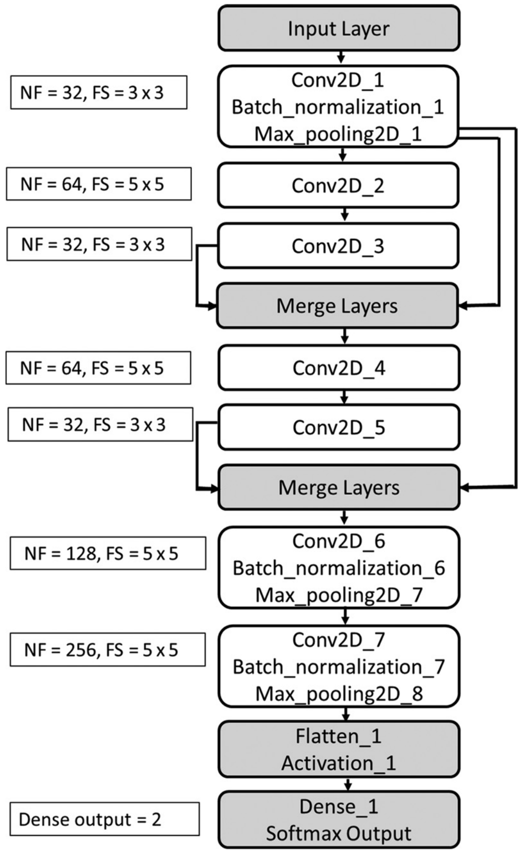

**FIG 2.** CNN architecture used in the experiment. Here NF represents number of filters and FS represents filter size.

### Classification Algorithm Testing and Technologist and Radiologist Survey

Forty-nine series not included in training (grayed-out cells of Fig 1) were set aside for DL classification testing. This dataset (of all orientations, contrast types, and pathologies) consisted predominantly of images with low/moderate levels of artifacts. Of these 49 series, 5 were initially rated by R0 as CB, 6 as CG and 38 as questionable. The same test series were also evaluated by 5 radiologists (D1–D5), with 3–18 years of experience in reading MR images, and 4 MR imaging technologists (T1–T4), with 5–26 years of experience in performing MR imaging. All 9 survey participants were asked to rate series as CG or CB, assuming that patients were scanned to rule out stroke and MS. Lower/higher IQ is typically required for stroke and MS, respectively. When rating, radiologists were not considering sequence appropriateness for diagnosis.

### Data Analysis

"Rescan" was defined as a series rated CG by the physician and CB by the technologist/DL. "Recall" was defined as a series rated CB by the physician and CG by technologist/DL. They represent false-positives and false-negatives, respectively. The true-positives and true-negatives (series called good/bad by both the radiologists and technologists) were not considered because they cause no additional burden to the health care system and require no corrective action. Differences between recall and rescan rates among different raters were analyzed using ANOVA in Minitab 12 (http://www.minitab.com/en-us/). Receiver operating characteristic (ROC) curves were computed in the Scikit-learn package (Python 2.7; https://scikit-learn.org/stable/index.html).

## RESULTS

### Rater Survey

Tables 1 and 2 summarize the number of series (of 49) rated as of insufficient quality by each rater and the unneeded rescans and recalls of different raters. It was assumed that patients were scanned to rule out MS (Table 1) and stroke (Table 2). The data in these tables highlight the fact that physicians differ in their tolerance for artifacts. For example, D1 can render a diagnosis when presented with a lower IQ than D2, D4, and D5. Additionally, radiologists' IQ ratings differed, depending on scan indication, in 36% of the cases surveyed, while technologists' IQ ratings only changed in 11% of the cases. Depending on who scans the patient and who reads the scan, there can be a large number of unneeded

**Table 1: Results of the survey—rule out MS clinical scan indication**[a]

| Doctor ID | No. Series of Insufficient Quality | Technician ID | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | T1 ($n$ = 26) | | T2 ($n$ = 31) | | T3 ($n$ = 12) | | T4 ($n$ = 13) | |
| | | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls |
| D0 | 24 | 7 | 5 | 7 | 0 | 2 | 14 | 0 | 11 |
| D1 | 13 | 14 | 1 | 19 | 1 | 4 | 5 | 2 | 2 |
| D2 | 35 | 4 | 13 | 4 | 8 | 0 | 23 | 0 | 22 |
| D3 | 24 | 11 | 8 | 13 | 5 | 2 | 13 | 0 | 10 |
| D4 | 28 | 7 | 8 | 6 | 2 | 1 | 16 | 0 | 14 |
| D5 | 30 | 4 | 8 | 4 | 3 | 1 | 19 | 0 | 17 |
| Mean ± SD | 25.7 ± 7.4 | 7.8 ± 4 | 7.2 ± 4 | 8.8 ± 6 | 3.2 ± 2.9 | 1.7 ± 1.4 | 15 ± 6.1 | 0.3 ± 0.8 | 12.7 ± 6.8 |

**Note:**—ID indicates identification. The numbers in parenthesis next to the technician identification numbers represent the total numbers of insufficient quality series identified by each rater.

[a] All numbers reported are from the 49 series of the survey. Each series was evaluated twice, assuming that the scan indication was MS and stroke.

**Table 2: Results of the survey—rule out stroke clinical scan indication[a]**

| Doctor ID | No. Series of Insufficient Quality | Technician ID | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | T1 (*n* = 12) | | T2 (*n* = 28) | | T3 (*n* = 7) | | T4 (*n* = 13) | |
| | | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls | Unneeded Rescans | Unneeded Recalls |
| D1 | 2 | 10 | 0 | 26 | 0 | 5 | 0 | 11 | 0 |
| D2 | 13 | 2 | 3 | 15 | 0 | 0 | 6 | 2 | 2 |
| D3 | 8 | 6 | 1 | 22 | 1 | 3 | 3 | 7 | 1 |
| D4 | 11 | 4 | 3 | 18 | 1 | 2 | 6 | 0 | 4 |
| D5 | 7 | 5 | 0 | 21 | 0 | 3 | 3 | 6 | 0 |
| Mean ± SD | 8.2 ± 4.2 | 5.4 ± 3 | 1.4 ± 0.7 | 20.4 ± 4.2 | 0.4 ± 0.2 | 2.6 ± 0.8 | 3.6 ± 1.1 | 6 ± 3.4 | 1 ± 0.4 |

**Note:**—ID indicates identification. The numbers in parenthesis next to the technician identification numbers represent the total numbers of insufficient quality series identified by each rater.

[a] All numbers reported are from the 49 series of the survey. Each series was evaluated twice, assuming that the scan indication was multiple sclerosis and stroke.

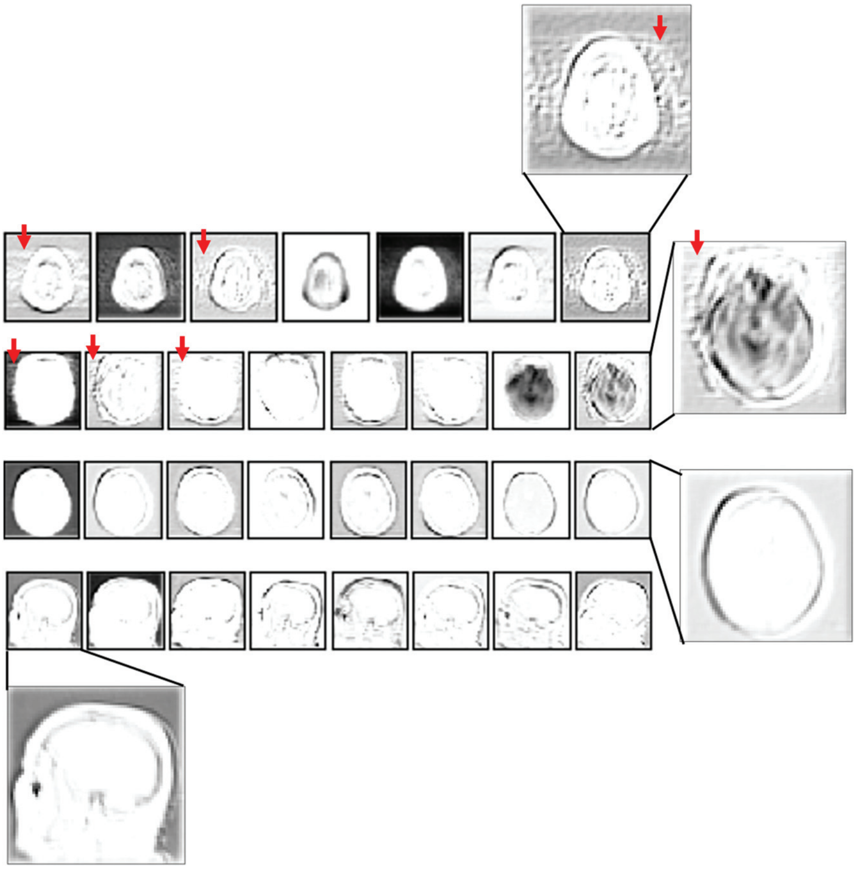

**FIG 3.** Representative filter responses from the fourth convolution layer of the CNN (Conv2D_4). *Rows 1 and 2*, Filter responses for motion-corrupted axial FLAIR/T2* input images, respectively. *Rows 3 and 4*, Filter responses from axial/sagittal T1 input images without motion, respectively. Filter responses are independent of image contrast and highlight the recognizable motion artifacts in the motion-corrupted images (*arrows*).

rescans or recalls. On average, technologists generated (4.7 ± 5.1)/(9.5 ± 6.8) rescans/recalls for the MS scan indications and (8.6 ± 7.7)/(1.6 ± 1.9) rescans/recalls for the stroke scan indication. Their IQ estimation for MS was generally underestimated (more recalls), and their IQ estimation for stroke was generally underestimated (more rescans).

### DL Classification: General Results

Representative convolutional filter responses from one of the intermediate steps of the CNN (Conv2D_4 layer of Fig 2) for different imaging contrasts (T1, FLAIR, and T2*) and in cases with/without motion are presented in Fig 3. In cases with motion (rows 1 and 2 of Fig 3), the filter responses show distinct patterns highlighting motion-related ghosting in the foreground and background of the response images (arrows). Filter responses are insensitive to image contrast, ensuring that a slightly different contrast, not seen during training, will likely not influence classification accuracy.

Three examples of DL classification results (using physician D0 as ground truth) are presented in Fig 4. While probabilities for individual slices can be variable, a few poor-quality slices do not necessarily imply that a rescan is needed. The rescan decision depends on the clinical scan indication and reading radiologist (Tables 1 and 2). Score prediction for a 25-slice volume took ~0.9 seconds. The ROC curve for the DL-based classification of the 49 test series (shown in On-line Fig 2) resulted in an area under the curve of 92%, comparable with the best classification results published elsewhere.[4]

### DL Classification: MS

Given the rating variability of Tables 1 and 2, it became clear that physicians tolerate different artifact levels. Consequently, the per-series score was checked against 3 different thresholds ($P[series] = [0.1, 0.5, 0.8]$) to decide the IQ rating of a given series. Table 3 summarizes the number of rescans and recalls of the DL with different thresholds, assuming that patients were scanned to rule out MS. Varying DL classification thresholds effectively change the rescan/recall ratio; lowering the threshold increases recalls, while increasing the threshold increases rescans. A single common threshold of 0.5 for all physicians results in (7 ± 4.7)/(5.3 ± 3.4) rescans/recalls, which represent a lower misclassification rate than that of the average technologist of (4.7 ± 5.1)/(9.5 ± 6.8), but without reaching statistical significance. Because of reading radiologists' varying artifact tolerances, a single threshold for highlighting series that need rescans is inefficient. Selecting an average threshold (eg, 0.5) results in too many rescans if an

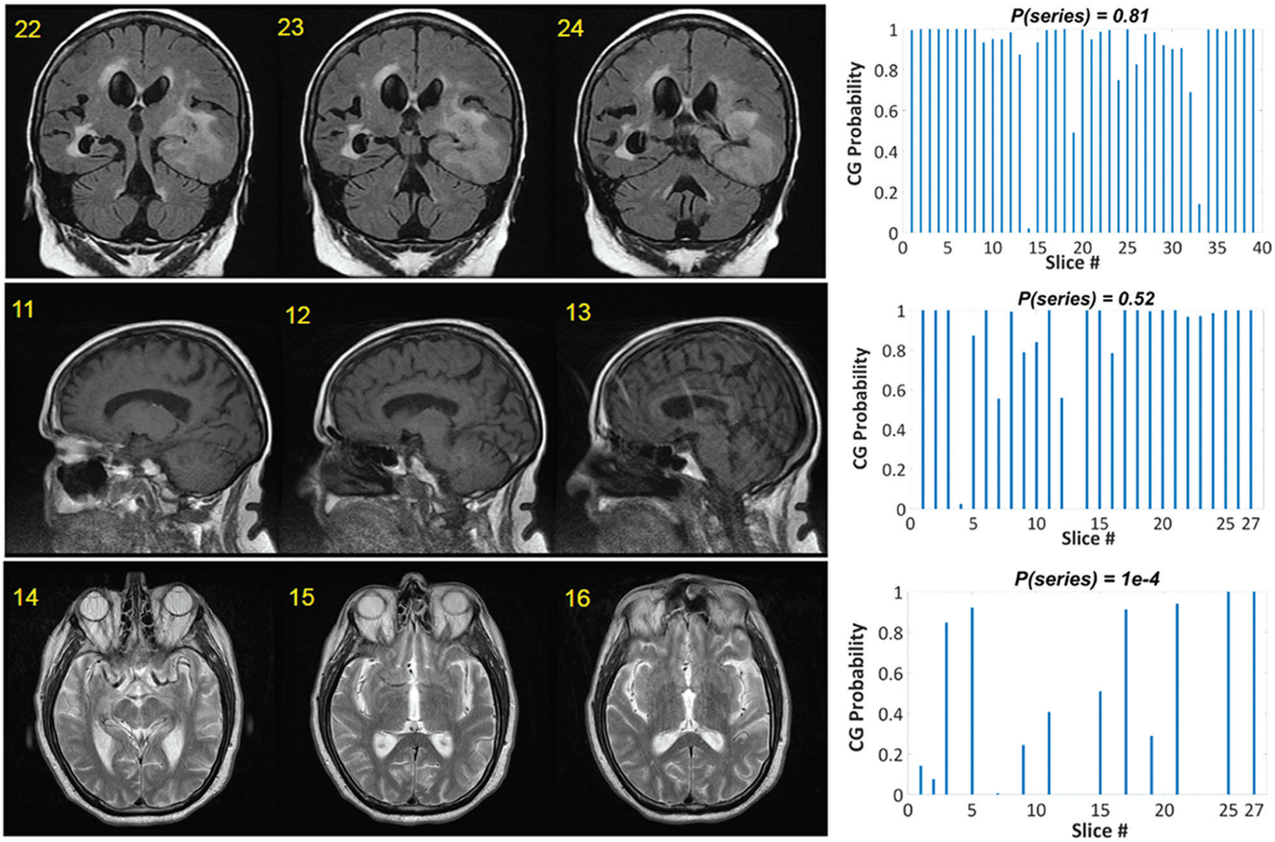

**FIG 4.** Examples of classification performance for 3 series. A few slices are displayed from each series (*left*), together with the slice ratings for the entire series (*right*). The numbers at the top left corner of each image represent the slice number.

**Table 3: Matrix documenting the number of unneeded rescans and recalls created by the DL approach with different thresholds, assuming that series were scanned to rule out MS**[a]

| | DL (T = 0.1) | | DL (T = 0.5) | | DL (T = 0.8) | |
|---|---|---|---|---|---|---|
| | Rescans | Recalls | Rescans | Recalls | Rescans | Recalls |
| D0 | 2 | 8 | 6 | 3 | 9 | 0 |
| D1 | 8 | 3 | 15 | 1 | 21 | 1 |
| D2 | 1 | 18 | 3 | 11 | 4 | 6 |
| D3 | 5 | 10 | 10 | 6 | 14 | 4 |
| D4 | 2 | 11 | 5 | 5 | 7 | 1 |
| D5 | 1 | 13 | 3 | 6 | 6 | 3 |
| Mean ± SD | 3.2 ± 2.8 | 10.5 ± 2.1 | 7 ± 4.7 | 5.3 ± 1.4 | 10.2 ± 2.6 | 2.5 ± 0.9 |

[a] All numbers are from the 49 test series. Here D0–D5 represent the same individuals as in Tables 1 and 2.

artifact-tolerant physician reads the scans (eg, D1), and in too many recalls, if an artifact-intolerant physician reads them (eg, D2).

Additionally, absolute minimization of the sum of rescans and recalls should not be the final goal of an automated rating approach. For example, a threshold of 0.1 for D3 results in the smallest sum of rescans and recalls. This, however, comes at the expense of generating 5/10 rescans/recalls, which may not be better than creating 10/6 rescans/recalls (which is what a threshold of 0.5 generates). One should aim for a constrained minimization of the sum of rescans and recalls, while maintaining an economically optimal rescan/recall ratio.

With automated rating tools, hospitals can enable per-radiologist thresholds resulting in optimal rescan/recall ratios. Such individualized thresholds could be easily implemented. For example, each radiologist can rate a single batch of datasets for a few clinical scan indications. For each indication, the concordance between the DL algorithm and the individual physician can be ascertained for a range of thresholds. The threshold resulting in the best DL-radiologist concordance will then be preserved for that physician for all ensuing scans. Assuming that a hospital aims for a rescan/recall ratio of 1.5 (equivalent to weighting rescan/recall classification errors by 0.4/0.6, respectively), the optimal thresholds for our 6 reading radiologists become 0.8/0.1/0.8/0.5/0.8/0.8, respectively. This results in 7.3 ± 2.2 rescans (statistically equivalent to the 4.7 ± 5.1 rescans caused by the technologists, $P = .22$) and 3.2 ± 2.5 recalls, which are statistically lower than the 9.5 ± 6.8 recalls caused by the technologists ($P = .03$).

### DL Classification: Stroke

The same DL network, without further training, was used to classify the same 49 series, assuming that patients were scanned to rule out stroke. While the probabilities output by the DL algorithm are independent of scan indication, it was hypothesized that adapting thresholds could compensate for the lower IQ needed from a scan obtained to rule out stroke. Lower thresholds of $P(Series) = (0.5, 0.1, 5e{-}4, 1e{-}6)$ were now explored. Table 4 summarizes the number of rescans and recalls of different raters, assuming that patients were scanned to rule out stroke. Significantly lower

**Table 4: Matrix documenting the number of unneeded rescans and recalls created by the DL approach with different thresholds, assuming that the series were scanned to rule out stroke[a]**

| | DL (T = 0.5) | | DL (T = 0.1) | | DL (T = 5*e*–4) | | DL (T = 1*e*–6) | |
|---|---|---|---|---|---|---|---|---|
| | Rescans | Recalls | Rescans | Recalls | Rescans | Recalls | Rescans | Recalls |
| D1 | 25 | 0 | 16 | 0 | 11 | 0 | 4 | 0 |
| D2 | 15 | 1 | 8 | 3 | 3 | 3 | 0 | 7 |
| D3 | 20 | 0 | 13 | 2 | 8 | 2 | 3 | 4 |
| D4 | 17 | 1 | 10 | 3 | 6 | 4 | 2 | 7 |
| D5 | 20 | 1 | 12 | 1 | 7 | 1 | 2 | 3 |
| Mean ± SD | 19.4 ± 3.8 | 0.4 ± 0.2 | 11.8 ± 1.4 | 1.8 ± 1.3 | 7 ± 2.9 | 2 ± 1.6 | 2.2 ± 1.5 | 4.2 ± 3 |

[a] All numbers are from the 49 test series. Here D1–D5 represent the same individuals as in Tables 1 and 2. Physician D0, whose ratings were used to train the DL algorithm, is now absent (as in Table 2) because no "stroke" ratings were available for this reader.

thresholds are now needed to separate CG from CB series in our DL-based approach. In fact, using the same average single classification threshold of 0.5, which was optimal for MS screening, would result in a generally higher misclassification rate than that of the average technologist ($P = .06$).

Assuming that the hospital aims for the same rescan-to-recall ratio of 1.5, the optimal thresholds for our 5 reading radiologists to rule out stroke become 1*e*–6/5*e*–4/1e–6/5*e*–4/1*e*–6, respectively. These result in (3.6 ± 1.5)/(2.8 ± 1.6) rescans/recalls, which are equivalent to the technologists' (8.6 ± 7.7)/(1.6 ± 1.9) rescans/recalls ($P > .05$).

## DISCUSSION

In this work, a fast, automated methodology to determine the diagnostic utility of a MRI series was demonstrated. A ResNet-inspired CNN architecture outputs a probability for each slice as belonging to CG or CB. A per-series score, computed as the geometric mean of the individual slice probabilities, is then compared to a threshold based on scan indication, or scan indication and reading radiologist, to decide whether a rescan is needed.

This approach was trained and tested on anatomic brain images of all orientations, contrasts, and pathologies for Cartesian sampling schemes. It was found that testing the architecture for the clinical scan indication on which it was trained (ie, MS) while using a single, cross-radiologist threshold of 0.5, results in fewer rescans/recalls (7 ± 4.7)/(5.3 ± 3.4) than the technologists' rescans/recalls of (4.7 ± 5.1)/(9.5 ± 6.8), without reaching statistical significance. With personalized thresholds, accounting for physicians' different artifact tolerances, this improvement became statistically significant, maintaining rescans to an equivalent (7.3 ± 2.2), but reducing recalls to (3.2 ± 2.5). The use of such an algorithm would closely match the radiologist assessing IQ in real-time.

When the same algorithm, without different training, was tested in rating the same image sets (now presumed to be acquired to rule out stroke), a significant lowering of the threshold was needed to render the algorithm's prediction similar to the physicians' rating. This finding is consistent with clinicians themselves requiring lower IQ images to diagnose stroke. Even with personalized thresholds, the rescans/recalls of (3.6 ± 1.5)/(2.8 ± 1.6) of the DL remained statistically equivalent to technologists' rescans/recalls of (8.6 ± 7.7)/(1.6 ± 1.9). This result is largely caused by the variability in the technologists' performances; a larger pooled technologist population may have altered the results.

This work is one of the first presenting evidence that MR imaging IQ is not an absolute measure, but a function of scan indication and reading radiologist. This information needs to be available for rating purposes to reduce rescans; otherwise, a single DL network with a single threshold could, in fact, increase the number of rescans or recalls. To maximize classification accuracy, a hospital can implement indication- and reading radiologist–level thresholds. Alternatively, considering that second-opinion radiologists and referring physicians may also read images, an indication-dependent threshold may be implemented that would work for the average physician, at the expense of decreasing classification accuracy.

While MRI examination indications span a broad range, grouping scan indications into 3–4 categories requiring similar IQ would likely suffice. For example, if 3 categories are chosen, the lowest acceptable IQ category could encompass scan indications such as screening for stroke, hemorrhage, or large masses. The second, midlevel IQ category could encompass screening for multiple sclerosis or spread of known tumor, while the third, highest IQ category could cover scan indications such as screening for epilepsy foci or small brain metastases.

Caution is advised when comparing our rescan/recall numbers with those in other studies. To test the performance of our classification algorithm in situations in which nonprofessional readers could not easily decide the clinical utility of a given series, our test data were purposely enriched in difficult cases; 78% of our test datasets were initially rated as questionable, which is higher than the occurrence of such cases in daily scanning. Consequently, our classification accuracy may appear artificially low.

Our per-slice rating approach is somewhat similar to one recently documented, in which the IQ of individual image patches was assessed using a 3-layer DL-based architecture.[13] Our full image classification, however, avoids false-positives in air-dominated patches. In addition to the architecture documented here containing residual in-network connections (Fig 2), a 5-convolutional-layer (no residual in-network connection) architecture was also tested. Our final implementation had a smaller size due to the reduced number of filters (8 versus 34 MB for the 5 convolutional layers) and a shorter prediction time (0.9 versus 1 second for the 5 convolutional layers in the same 25 slices). With identical, full training, the 2 implementations had comparable classification accuracy. While only using one-third of the data for training, our final implementation outperformed the 5-convolutional-layer architecture (area under the curve for D0 of 86% versus 92% in our 49 test datasets).

To understand the potential economic impact of automated IQ rating, we used the assumptions of Andre et al.[1] Without published recall rates or institutional costs, recall rates at 1 of the authors' outpatient imaging facilities, Albany Medical College, were first surveyed. Among the recalled examinations (0.6%), most were due to lack of contrast uptake, incorrect protocol, or scanner failure. Only 6 examinations in 1 calendar year (0.05%) were due to patient motion. At $600/brain examination,[1] the 6 recalls caused $3600 in revenue loss. Scaling this number up to the outpatient/inpatient proportion of Andre et al[1] and considering that inpatients/outpatients cause rates of 7.5%/29.4% of moderate and severe artifacts,[1] a recall-induced revenue loss of $7700 results. Assuming that our test series are reflective of the examinations performed at a given site in 1 year, technologists generated 4.7/9.5 rescans/recalls (MS) and 8.6/1.6 rescans/recalls (stroke), for a total of 13.3 rescans, costing $115,000,[1] and 11.1 recalls, costing $7700 (as per the calculation above). Our test series enrichment in questionable datasets has no impact on this economic calculation because everything is scaled by the documented loss/site. By using individualized thresholds for clinical scan indications and reading radiologist, DL generated 7.3/3.2 rescans/recalls (MS), and 3.6/2.8 rescans/recalls for stroke, for a total of 10.9 rescans and 6 recalls. Scaling these numbers by the cost of rescans/recalls per site results in a revenue loss of $98,400 with DL versus $122,700 without DL. More than $24,000 savings per site are obtained without negatively affecting patient care.

Interestingly, it was found that, although technologists generate comparable rates of rescans and recalls, the rate of motion-induced examination recalls was remarkably low (0.05%). Recalls are aggressively avoided because they are costly and are not covered by insurance. Patients who moved during MR imaging examinations are sometimes directed to follow-up contrast CT. Such examinations are shorter and hence have higher compliance rates. They are also generally reimbursed by insurance, hence not affecting the profits of the imaging center. Second, there is some redundancy in the prescribed series, and radiologists can often perform a diagnosis with only a few series of diagnostic quality. While this fact suggests that an examination-level (as opposed to a series-level) automated rating may be more appropriate, such examination-level ratings would not be actionable, as they would not be able to highlight the series that needs to be rescanned.

This study has a few limitations. Only single-site, 1.5T data were used for training and testing, while DWI was excluded. Further performance validation will need to include 3T data. In addition, only brain data were used for both training and testing; while no brain-specific features were extracted, it remains to be tested how well this automated method works with other anatomies and multiple *k*-space sampling schemes.

## CONCLUSIONS

A fast, deep learning based approach similar to the one described here could soon aid technologists in deciding whether a MRI series needs to be rescanned, thereby reducing rescan and recall rates. For optimal performance, scan indication or scan indication and reading radiologist information will need to be provided to the algorithm.

Disclosures: Arathi Sreekumari—*UNRELATED*: *Employment*: GE Healthcare. Dattesh Shanbhag—*UNRELATED*: *Employment*: GE Healthcare. Desmond Yeo—*UNRELATED*: *Employment*: GE Healthcare; *Grants/Grants Pending*: National Institutes of Health, *Comments*: National Institutes of Health notice of award pending for an R01 for which I am a site Principal Investigator*; *Stock/Stock Options*: GE Healthcare; *Other*: University of Michigan, *Comments*: I am an external advisory board (EAB) member of the University of Michigan Biomedical Engineering Department. Aside from receiving lodging from the University of Michigan for annual EAB meetings, no financial remuneration is provided for this role. Thomas Foo—*UNRELATED*: *Employment*: GE Global Research. Uday Patil—*UNRELATED*: *Consultancy*: GE India Industrial Private Limited (Division: John F Welch Technology Centre, formerly GE India Technology Centre Pvt Ltd). John Port—*UNRELATED*: *Consultancy*: Biomedical Systems, Neuronetics Inc, *Comments*: I serve as an image-reading service for both of these companies. Money paid to Mayo Clinic*; *Patents (Planned, Pending or Issued)*: Mayo Clinic, *Comments*: I have a provisional patent under consideration.* Ileana Hancu—*UNRELATED*: *Employment*: GE Healthcare. Julie Pilitsis—*UNRELATED*: *Grant*: GE Global Research.* *Money paid to the institution.

## REFERENCES

1. Andre JB, Bresnahan BW, Mossa-Basha M, et al. **Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical MR examinations.** *J Am Coll Radiol* 2015;12:689–95 CrossRef Medline
2. Pizarro RA, Cheng X, Barnett A, et al. **Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning 28066227 algorithm.** *Front Neuroinform* 2016;10:52 95 CrossRef Medline
3. Hagens MH, Burggraaff J, Kilsdonk ID, et al. **Impact of 3 Tesla MRI on interobserver agreement in clinically isolated syndrome: a MAGNIMS multicentre study.** *Mult Scler* 2018 Jan 1. [Epub ahead of print] CrossRef Medline
4. Gatidis S, Liebgott A, Schwartz M, et al. **Automated reference-free assessment of MR image quality using an active learning approach: comparison of support vector machine versus deep neural network classification.** In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine,* Honolulu, Hawaii. April 22–27, 2017; 3979
5. Elsaid N, Roys S, Stone M, et al. **Phase-based motion detection for diffusion magnetic resonance imaging.** In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine,* Honolulu, Hawaii. April 22–27, 2017; 1288
6. Kelly C, Pietsch M, Counsell S, et al. **Transfer learning and convolutional neural net fusion for motion artefact detection.** In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine,* Honolulu, Hawaii. April 22–27, 2017; 3523
7. Hirsch J, Kohn A, Hoinkiss D, et al. **Quality assessment in the multicenter MR imaging study of the German national cohort.** In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine,* Honolulu, Hawaii. April 22–27, 2017; 4458
8. Guehring J, Weale P, Zuehlsdorff S, inventors. System for dynamically improving medical image acquisition quality. US patent US 8,520,920 B2, 2010
9. Blumenthal JD, Zijdenbos A, Molloy E, et al. **Motion artifact in magnetic resonance imaging: implications for automated analysis.** *Neuroimage* 2002;16:89–92 CrossRef Medline
10. Kelly C. **Transfer learning and convolutional neural net fusion for motion artifact detection.** *In: Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine,* Honolulu, Hawaii. April 22–27, 2017*;* 3523
11. Lorch B, Vaillant G, Baumgartner C, et al. **Automated detection of motion artefacts in MR imaging using decision forests.** *J Med Eng* 2017;2017:4501647 CrossRef Medline
12. Meding K, Lokyushin A, Hirsch M. **Automatic detection of motion artifacts in MR images using CNNS.** In: *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing,* New Orleans, Louisiana. March 5–9, 2017
13. Küstner T, Liebgott A, Mauch L, et al. **Automated reference-free detection of motion artifacts in magnetic resonance images.** *MAGMA* 2018;31:243–56 CrossRef Medline
14. Esses S, Lu X, Zhao T, et al. **Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture.** *J Magn Reson Imaging* 2018;47:723–28 CrossRef Medline
15. Chollet F. Keras Documentation. 2017 https://keras.io/ Accessed December 5, 2018
16. He K, Zhang X, Ren S, et al. **Deep residual learning for image recognition.** In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, Nevada. June 26 to July 1, 2016
17. Karpathy A. Convolutional Neural Networks for Visual Recognition. 2018 http://cs231n.github.io/ Accessed December 9, 2018