TITLE: BrainQCNet: a Deep Learning attention-based model for multi-scale detection of artifacts in brain structural MRI scans

AUTHORS: Mélanie Garcia[1,3], Nico Dosenbach[4], Clare Kelly[1,2,3]

AFFILIATIONS: [1] Department of Psychiatry at the School of Medicine, Trinity College Dublin, Dublin, Ireland, [2] School of Psychology, Trinity College Dublin, Dublin, Ireland, [3]Trinity College Institute of Neuroscience, Trinity College, Dublin, Ireland, [4]Department of Neurology, Washington University School of Medicine, St. Louis, Missouri 63110, USA

KEYWORDS: quality control, QC, structural MRI, Deep Learning, interpretable

CORRESPONDING AUTHOR:
Clare Kelly,
Trinity College Institute of Neuroscience,
Trinity College,
Dublin 2,
Ireland
clare.kelly@tcd.ie

Abstract

Analyses of structural MRI (sMRI) data depend on robust upstream data quality control (QC). It is also crucial that researchers retain the maximum amount of usable data to ensure reproducible, generalisable models. The time-consuming task of manual QC evaluation has prompted the development of tools for the automatic assessment of brain sMRI scans. Such tools are particularly valuable in this age of big data. One limitation of the most commonly used tools is that execution time is long, which poses a challenge in terms of duration and resource usage, particularly when processing large datasets. Further, evaluation is global (pass/fail) rather than localized. Having a tool that localizes areas of low quality could prevent unnecessary data loss. To address these issues, we trained a Deep Learning model, ProtoPNet, to classify minimally preprocessed 2D slices of scans that were manually annotated with a refined quality assessment (ABIDE 1 *n* = 980 scans). To validate the best model, we assessed 2141 ABCD scans for which gold-standard manual QC annotations were available. We obtained excellent accuracy: 82.4% for good quality scans (Pass), 91.4% for medium to low quality scans (Fail). Further validation using 799 scans from ABIDE 2 and 751 scans from ADHD-200 confirmed the reliability of our model. Accuracy was comparable to or exceeded that of another commonly used tool (MRIQC), but with dramatically reduced processing and prediction time (1 min per scan, GPU machine, CUDA-compatible). To facilitate faster and more accurate QC prediction for the neuroimaging community, we have shared the model that returned the most reliable global quality scores, local predictions of quality, and maps and prototypes of local artifacts as a BIDS-app (https://github.com/garciaml/BrainQCNet).

Abbreviations:

- **CNN**: Convolutional Neural Networks, a category of Deep Learning algorithm
- **ML**: Machine Learning
- **DL**: Deep Learning
- **ProtoPNet**: Prototypical Part Network model
- **VGG19**: Visual Geometry Group model, a type of very deep convolutional neural network with 19 layers in the model;
- **ResNet152**: Residual Networks model with 152 layers
- **DenseNet161**: Densely Connected Convolutional Networks with 161 layers
- **proto-VGG19:** ProtoPNet model with a VGG19 architecture in the CNN part
- **proto-ResNet152:** ProtoPNet model with a ResNet152 architecture in the CNN part
- **proto-DenseNet161:** ProtoPNet model with a DenseNet161 architecture in the CNN part

1. Introduction

Analyses of structural MRI (sMRI) data depend on robust upstream data quality control. This is particularly true for predictive analyses incorporating machine learning techniques, where artifacts and noise may severely bias results and jeopardize generalisability (Backhausen et al., 2016; Gilmore et al., 2019; White et al., 2018; Reuter et al., 2015). Artifacts related to participant motion are a particular concern when working with very young participants, or those with neurodevelopmental diagnoses, such as Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder (Rauch, 2005; Nordahl et al., 2016). In such settings, data collection is usually a demanding and costly task, and it is crucial that researchers retain the maximum amount of usable data to build realistic models.

In this age of big data, manual QC evaluation of sMRI data through visual inspection is a time-consuming and monotonous task, prompting the development of new tools for automatic (full or partial) quality assessment of brain sMRI scans (Esteban et al., 2017; Sujit et al., 2019; Zarrar et al., 2015; Keshavan et al., 2019; White et al., 2018; Alfaro-Almagro et al., 2018; Glasser et al., 2016; Marcus et al., 2013). Generally, these tools compute a number of diagnostic metrics using sMRI data to help researchers sort images prior to any analysis. One such tool, MRIQC (Esteban et al., 2017), has revolutionized QC of MRI data by providing a reliable and accurate Machine Learning-based assessment of scan quality that has been made freely available to the neuroimaging community as an open-source application (https://mriqc.readthedocs.io/en/stable/). The tool extracts 64 image quality metrics that were chosen on the basis of the Preprocessed Connectomes Project (PCP) Quality Assessment Protocol (Zarrar et al., 2015) and include measures such as Contrast to Noise Ratio and Entropy Focus Criterion (Esteban et al., 2017). The MRIQC algorithm uses Machine Learning to find a function that predicts a global quality score for each scan using these metrics. Although highly accessible, automated, and accurate, growth in the size of datasets (e.g., thousands to tens of thousands of sMRI scans for database such as ABCD (Volkow et al., 2018; Karcher and Barch, 2021), ENIGMA (Whelan et al., 2018) and UK Biobank (Sudlow et al., 2015)) and increasing concern about energy usage prompted us to investigate whether there was scope to build on the progress of MRIQC to further advance automated QC.

We identified two primary opportunities for development. First - the time and resources required to assess each sMRI scan. Because the MRIQC prediction is based on a large number of image quality metrics (64) computed for each scan, it is relatively demanding in terms of time (~45 minutes per scan), and by consequence, energy resources. Although some of this image processing may be exploited in subsequent analyses, extracting these metrics for all scans means that processing resources are expended on scans that are ultimately unusable due to poor quality. When working with very large databases (>1000 scans), MRIQC may take a long time to complete, unless computations can be parallelized on High Performance Clusters. Second, the quality score returned by MRIQC is a global one. For some scans, areas of low quality, artifact, or corruption may be circumscribed; uncorrupted areas might still be of interest for certain studies (e.g., focused on subcortical regions or cerebellum rather than cortex). A quality assessment that included both global and local quality assessments would minimize data loss.

Deep Learning algorithms have the potential to address these two issues. While training a Deep Learning model may initially take longer than a traditional Machine Learning (ML) algorithm (because there are more parameters to train), the subsequent processing and inference time is reduced compared to ML, thanks to the chain of simple computations performed, particularly in the context of image processing and on GPU machines. This rapid inference makes DL models more scalable for Big Data applications. In addition, it has been shown that Convolutional Neural Networks (CNN) - a category of Deep Learning algorithms - can process images more efficiently than traditional image processing methods, by considerably reducing processing time while generally increasing accuracy (Hastie et al., 2009; LeCun et al., 1999).

Yet, the medical imaging community has been wary of CNN, possibly due to their more complex and abstract nature, which leads to difficulties with interpretability. Recent improvements in the interpretability and clinical utility of such models may address these concerns. One such development is the use of visual attention models. These models mimic human visual attention by focusing on the relevant parts of an image in the task of image recognition. For example, when recognising a bird in an image, a person might look at different levels of detail in the image, such as the size, the color, shape of the beak, etc. Attention-based algorithms mimic this process through different mathematical and implementation designs. These models expose the parts of an input the network algorithm focuses on (identifies as most strongly predictive). For instance, class activation mapping (Zhou et al., 2016) provides an interpretation at the object level (in our example, a map with an activated area covering the bird) while other models provides an interpretation at different parts of the image (in our example, several maps with activated areas covering the beak only, a specific color on the bird, etc.) (Chen et al., 2019; Zhang et al., 2014; Zheng et al., 2017). ProtoPNet is a CNN algorithm that provides this kind of refined part-level interpretation in addition to another level of interpretability: it points to prototypical cases that are similar to the parts identified as predicted (i.e., focused on).

MRI studies have started to integrate the attentional approach within known Deep Learning models, such as the segmentation algorithm U-Net combined with an attention mechanism (Khanh et al., 2020) and brain tumor detection (Ranjbarzadeh et al., 2021). Here, we leveraged the advantages of Deep Learning models with attention mechanisms to perform automated QC of sMRI data. We trained an attention model to perform QC assessments of minimally processed developmental sMRI data, including data collected from participants with neurodevelopmental diagnoses. Specifically, we trained the CNN ProtoPNet, as described above (Chen et al., 2019). The process used by the algorithm is similar to the one humans use when we perform manual classification of MRI scans. That is, we visually search for the presence of artifacts, slice by slice, in 2D. To recognise and distinguish the types of artifacts on a scan, we compare the slice to slices from other scans that have similar flaws. ProtoPNet imitates this human attention process artificially. A key advantage of this model is that it can return local quality scores for every pixel of a 2D slice of a 3D scan, along with a global quality score.

Among the different layers of ProtoPNet, the model has a Convolutional layer corresponding to a CNN, which can be pre-trained on appropriate data (here, MRI images). We compared three different pre-trained CNN models: VGG19 (Simonyan and Zisserman, 2015), ResNet152 (He et al., 2015) and DenseNet161 (Huang et al., 2018). To train our algorithms, we used 980 structural brain MRI scans from the ABIDE 1 dataset (Di Martino et al., 2014). We validated the best model using the gold standard test: independent, multisite data. Specifically, we validated the best model using 2141 scans from ABCD (Volkow et al., 2018; Karcher and Barch, 2021), 799 scans from ABIDE 2 (Di Martino et al., 2017) and 751 scans from ADHD-200 (Bellec et al., 2017). A key advantage of our algorithm over existing approaches is that it requires only minimal preprocessing, which dramatically reduces the total processing time for every scan (1 minute on a GPU machine, 20 minutes on a CPU machine). In the context of the growing use of enormous datasets containing tens of thousands or even tens of thousands of participants, our method could offer substantial savings in terms of time and computational resources. Across our independent validation datasets, we show excellent accuracy that matches or surpasses existing automated QC algorithms.

2. Materials and Methods

2.1 Datasets and pipeline summary

In our study, we used structural MRI data from ABIDE 1 (Di Martino et al., 2014), ABIDE 2 (Di Martino et al., 2017), ADHD-200 (Bellec et al., 2017) and ABCD (Volkow et al., 2018; Karcher and Barch, 2021). Details of each of the datasets used are provided in **Table 1**.

	N Scans	QC metrics	Age	Gender	*N* Sites
ABIDE 1	980	PCP [11] metrics computed;Ternary manual annotation by 3 judges judgment : “OK”, “maybe”, “fail”.	min = 6.5 q_25% = 11.6 median = 14.7 q_75% = 20 max = 64	F: 147 M: 833	20
ABIDE 2	799	None available	min = 5.1 q_25% = 9 median = 11 q_75% = 14.7 max = 64	F: 202 M: 597	14
ADHD-200	751	Binary manual annotation: 0 for pass; 1 for fail.	min = 7 q_25% = 9.2 median = 11 q_75% = 13.6 max = 21.8	F: 326 M: 424 1 unknown	7
ABCD	2141	Ternary manual annotation: pass, questionable, fail	min = 0 q_25% = 6 median = 12 q_75% = 23 max = 81	F: 1153 M: 988	unknown

Table 1. Dataset descriptions.

A schematic of our study pipeline is shown in Section 9.2 (Supplemental Information). A summary of the process is as follows:

1. We performed detailed manual QC (Backhausen et al., 2016) of 980 scans from ABIDE 1 database (Di Martino et al., 2014). Although pass/fail/maybe annotations are provided with the dataset (see Table 1), our manual annotation captured additional detail about the different types of artifact present on a given scan.
2. We built a training set and a validation set that served to train a Deep Learning algorithm in the task of the detection of structural MRI artifacts. The Deep Learning algorithm we used is called ProtoPNet.
3. We used three different model architectures (VGG19, ReNet152[f], DenseNet161) and trained the ProtoPNet algorithm using different numbers of epochs. An epoch is a single step within which the algorithm has been optimized by all the images of the training set. Because of GPU memory issues, optimization is achieved through an iterative process: we optimize the algorithm with batches of data of size *n*, which is smaller than the full size of the training set, *N*.
4. We selected the best model on the basis of ROC AUC and accuracy scores on the training and validation sets, and on the first testing set (908 scans from ABIDE 1 (Di Martino et al., 2014))
5. We validated the best model on three independent testing sets (799 scans from ABIDE 2 (Di Martino et al., 2017), 751 scans from (Bellec et al., 2017), 2141 scans from ABCD (Volkow et al., 2018; Karcher and Barch, 2021)).

Importantly, prior to the steps involved in converting the 3D scans to 2D slices, and data augmentation, no preprocessing was applied to the sMRI scans.

2.2 Manual Quality Control Annotation

Inspired by the work of (Backhausen et al., 2016), we manually annotated MRI scans from ABIDE 1 according to a classification scheme specifying four different types of artifacts: (1) blurring (global or local), (2) ringing, (3) low contrast noise ratio (CNR) of subcortical structures, (4) low contrast noise ratio between grey matter and white matter. For each slice of each 3D scan, we also noted whether each observed artifact was visible locally or globally on the 2D slice, and on what axis (sagittal, coronal, axial). When no artifact was observed, we labeled the 3D scan as “good quality” (Class 0). Otherwise, we labeled the 3D scan as being corrupted (Class 1; see **Figure 1**), keeping in mind that Class 1 is a wide spectrum that includes scans with localized artifacts as well as very low quality, globally disrupted scans.

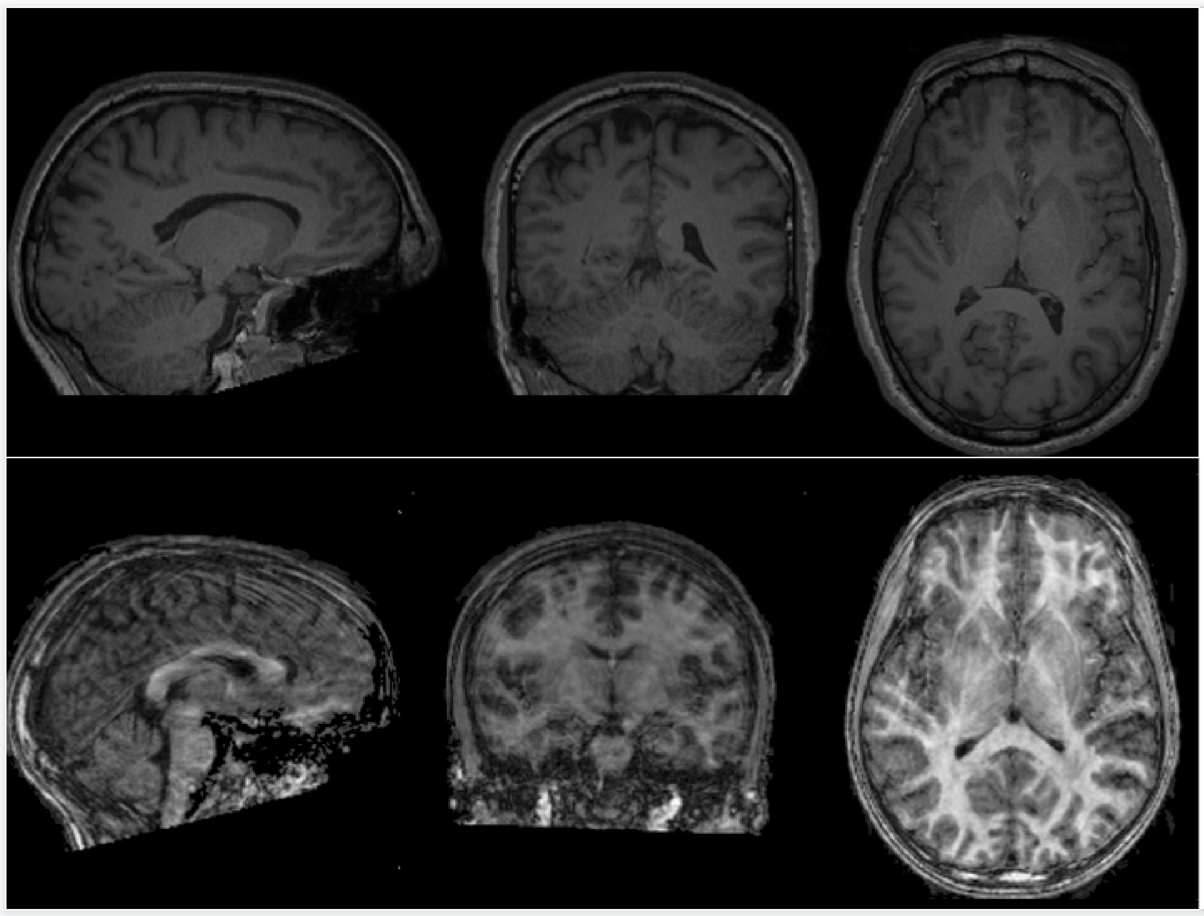


Figure 1. Example of a good quality scan (top panel - Class 0) and a very low quality scan (lower panel - Class 1)

2.3 Training & Validation sets

We built an initial set of images on which to train our Deep Learning algorithm from 30 highly corrupted/distorted scans (Class 1) and 30 high quality scans randomly selected from Class 0. We validated every 2 epochs by assessing the prediction accuracy of the model for 6 additional very low quality scans (i.e., scans with clearly identifiable global artifact/corruption) and 6 high quality scans. Highly corrupted scans were included in both the training and validation sets in order to maximize the chances of obtaining meaningful prototypes representative of scan artifacts and corruption.

Chen et al. (2019) showed that the ProtoPNet network algorithm worked better on cropped images, so each 3D scan was cropped to remove black areas, then converted from Nifti format to 2D PNG images (using Med2Image https://github.com/FNNDSC/med2image). For each scan there were between 150-200 2D slices in every 3 directions - sagittal, coronal, axial - approximately 450-600 images per scan. The first and last 20 slices of each resulting image stack were discarded since they contained little brain tissue or artifacts. Taking a random sample of 50 slices per axis per scan, we then created a training set comprising 4500 very low quality and 4500 good quality slices from all the 60 participants of the training set, and a validation set of 1800 slices, also balanced for quality.

Next, this training set was augmented with a set of random transformations (using the library Augmentor https://github.com/mdbloice/Augmentor) which rotated, skewed, and sheared the images. This yielded an augmented training set of 270000 images. Data augmentation is used to prevent overfitting in Deep Learning, thus improving generalizability of the algorithms.

All 2D images from good quality scans were defined as Label 0, and all 2D images from low quality scans were defined as Label 1. The algorithm was trained to perform a binary classification between Label 0 and Label 1 images using the augmented training set, and validation accuracy was computed every 2 epochs.

2.4 Deep Learning Algorithm

The algorithm we used - ProtoPNet (Chen et al., 2019) - is a Deep Learning Attention model that reproduces the human manual process for classifying images.

The network consists of a regular convolutional neural network, followed by a prototype layer and a fully connected layer with weight matrix and no bias. In our experiment we used three different architectures for the regular convolutional network:VGG19 (Simonyan and Zisserman, 2015), ResNet152 (He et al., 2015) and DenseNet161 (Huang et al., 2018). These three models are well known Deep Learning algorithms for image classification. They have shown great performance in 2D [6-8]. We compared these three models integrated in the ProtoPNet model because they are all performant algorithms with different architectures, leading to variable benefit on the number of parameters, the capacity to fit the data, etc. More globally in Machine Learning, it is appropriate to compare different types of algorithm for a same problem, to detect overfitting and to retain the best type of algorithm for the given problem [25].

In their approach, Chen et al. (2019) constrained each convolutional filter to be identical to some latent training patch, in order to make every convolutional filter interpretable as visualisable prototypical image parts. In our study, the “prototypes” or “prototypical images” corresponded to the Class 0 (good quality) and Class 1 (poor quality) images of the augmented training set. The algorithm works, in part, by comparing images in the validation and test sets to parts of the prototypes. The number of images selected randomly as prototypes during each epoch of training was set to 2000.

In the ProtoPNet global architecture, the prototype layer computes similarity scores between the convolutional filters of the input image and the ones from the 2000 prototypes at a fixed epoch.The similarity scores are computed with an inverted L2 norm distance.

Chen et al. [5] explained that given a convolutional output $z = f(x)$, the j-th prototype unit g_{p_j} in the prototype layer g_p computes the squared L^2 distances between the j-th prototype p_j and all patches of z that have the same shape as p_j, and inverts the distances into similarity scores. The result is an activation map of similarity scores whose value indicates how strong a prototypical part is present in the image [5].

Mathematically, the prototype unit g_{p_j} computes

$g_{p_j}(z) = max_{\tilde{z} \epsilon patches(z)} log((||\tilde{z} - p_j||_2^2 + 1)/(||\tilde{z} - p_j||_2^2 + \epsilon))$. The function g_{p_j} is monotonically decreasing with respect to $||\tilde{z} - p_j||_2$ (if $\tilde{z}$ is the closest latent patch to p_j) [5]. If the output of the j-th prototype unit g_{p_j} is large, then there is a patch in the convolutional output that is (in 2-norm) very close to the j-th prototype in the latent space, and this in turn means that there is a patch in the input image that has a similar concept to what the j-th prototype represents [5].

Next, the fully connected layer predicts the class of the input image from the 2000 similarity scores. We obtained the probability scores by applying the softmax function to the output logits of the fully connected layer. In theory, this method of regularization and comparison should improve the generalizability of the algorithm. That is why, despite a small training set, we expected the algorithm to deliver good results. More mathematical details of the ProtoPNet model are given in (**Chen et al., 2019)**, and **Figure 2** illustrates its architecture in our context.

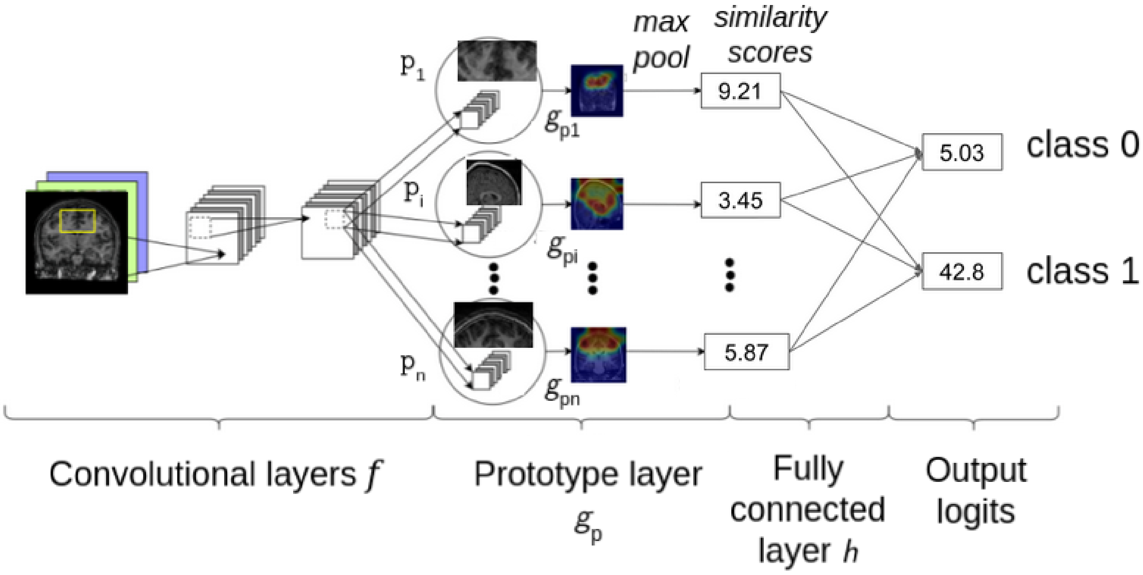


Figure 2. Architecture of the model; example for a very low quality scan.

We initiated the training with pre-trained models - VGG19, ResNet152, DenseNet161 - on ImageNet (Deng et al., 2009), drawn from the model zoo of Pytorch (https://pytorch.org/serve/model_zoo.html). We used the same initialisation parameters as previous experiments (Chen et al., 2019), including 5 “warming” epochs for which no accuracy was computed. As a reminder, each epoch is a step during which the algorithm is optimized by all the images of the training set. Because of the GPU memory demands of this process, optimization is achieved iteratively: we optimize the algorithm using small batches of data. Here, we used the same batch sizes as the original study by Chen et al. (2019) - 80 for the training and 100 for the testing phase.

We trained our models in a distributed way on AWS cloud instances of type p3.8xlarge and p3.16xlarge initialized with the AMI Deep Learning. The instances correspond to 4 or 8 GPUs NVIDIA V100. We trained ResNet152 on 20 epochs and VGG19 and DenseNet161 on 30 epochs. We saved models and associated prototypes every 10 epochs.

Finally, we integrated the best model to an open-source BIDS-app (Gorgolewski et al., 2017) we developed, to share it with the neuroimaging community in a ready-to-use format. BIDS (Brain Imaging Data Structure; Gorgolewski et al., 2016) is a community effort aimed at providing a standardized way of organizing neuroscience datasets that has facilitated the development of a number of open source analysis pipelines and applications. Instructions to use our app are available here: https://github.com/garciaml/BrainQCNet.

2.5 Independent Validation Sets

After training the models, we performed a validation on separate testing sets that consisted of all the slices from 4599 full 3D brain sMRI scans:

- 908 scans from ABIDE 1 (Di Martino et al., 2014) that were used to select the best model;
- 2141 scans from ABCD (Volkow et al., 2018; Karcher and Barch, 2021) that were used to validate the best model;
- 799 scans from ABIDE 2 (Di Martino et al., 2017) that were used to validate the best model (see Section 9.3; Supplemental Information);
- 751 scans from ADHD-200 (Bellec et al., 2017) that were used to validate the best model (see Section 9.3; Supplemental Information).

2.6 MRIQC

MRIQC (Esteban et al., 2017) is currently the reference algorithm for assessing automatically the quality of brain structural and functional MRI scans. It is based on a Machine Learning algorithm that was trained on a large number of metrics of quality previously extracted and computed from raw scans. As outlined in the introduction, these metrics were chosen as part of the Preprocessed Connectomes Project (PCP) Quality Assessment Protocol (Zarrar et al., 2015) to harmonise the assessment of the quality of brain MRI scans (Zarrar et al., 2015), like the signal-to-noise ratio. The output of MRIQC is a score and a binary prediction pass/fail outcome for each scan.

The main disadvantage of MRIQC is that it takes about 45 minutes to compute a QC result, mainly because of all the preprocessing steps involved in extracting the quality metrics.

Nevertheless, since this method is reliable (accuracy estimated to 76%±13% on new sites, using leave-one-site-out cross-validation, accuracy of 76% on a held-out dataset of 265 scans; Esteban et al., 2017), and widely employed, we used it here to generate predictions of the quality of each scan on ABIDE 2 (Di Martino et al., 2017; 799 scans). We treated these MRIQC-based predictions as the “ground truth” with which we compared the results of our algorithm.

We also compared the distribution of the scores returned by MRIQC for ABIDE 1 (Di Martino et al., 2014) (980 scans) with the distribution of scores returned by our models. In particular, we analized the discrimination between good quality scans and medium quality and low quality ones.

2.7 Data Ethics statement

The three databases used in the project - ABIDE 1, ABIDE 2, ADHD200 - are shared by the International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/). Each dataset was fully de-identified and anonymized in accordance with the US Health Insurance Portability and Accountability Act (HIPAA). All the datasets were collected and shared in accordance with the local regulations on ethics and data protection. Data usage is unrestricted for non-commercial research purposes; it is openly shared with the scientific community under the license Creative Commons BY-NC-SA. Our work with these open data is approved by the Research Ethics Committee of the School of Psychology at Trinity College Dublin.

Data from the ABCD study was fully de-identified and anonymized, and each data-collecting site obtained informed consent from participants and their parents/guardians. The ABCD study developed guidelines for ethical considerations to be applied by each data-collecting site, and organized a hierarchy of workgroups who assessed whether each step of the collection process conformed to the ABCD guidelines (Clark et al., 2018).

2.8 Materials and code availability

The three databases used in the project - ABIDE 1, ABIDE 2, ADHD200 - are openly shared by the International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/). The ABCD database is available upon request (https://nda.nih.gov/abcd/request-access).

All information about how sample size and data exclusion was determined, inclusion criteria (established prior to data analysis), and all derived measures used in this study are described in the Methods and Results sections. No part of the analysis was pre-registered prior to the research being conducted.

To maximize the reproducibility of our analyses and usability of our model, all code for preprocessing and predicting the quality of structural MRI data scans ordered is available as a BIDS-database in https://github.com/garciaml/BrainQCNet_CPU for users of CPU machines and in https://github.com/garciaml/BrainQCNet_GPU for users of GPU machines compatible with CUDA technology. Documentation for our BIDS-app on CPU or CPU is available here: https://github.com/garciaml/BrainQCNet.

All global predictions of quality for the 4671 scans we used from the ABIDE 1 & 2, ADHD200 and ABCD databases are available through the GitHub repository: https://github.com/garciaml/BrainQCNet.

3. Results

3.1 Annotations

Manual QC inspection of 980 scans from ABIDE 1 (Di Martino et al., 2014) identified 564 high quality scans, 36 very low quality scans (that we used in the training and validation sets), and 380 scans with either local artifacts or with mild-moderate global corruption (used in the testing set).

Local ringing (likely reflecting motion) was the most commonly occurring local artifact, and was often combined with other artifact types.

3.2 Training performance

In the results and figures below, we use the following naming convention: the prefix “proto-” corresponds to the ProtoPNet algorithm, while the suffix indicates the CNN architecture: VGG19, ResNet152, or DenseNet161 (see subsection 2.4 of section 2. Materials and Methods).

We obtained excellent accuracy for the detection of good (Class 0) and bad (Class 1) quality slices during training. From epoch 10, the accuracies of the three models - proto-VGG19, proto-ResNet152, proto-DenseNet161, were above 99% on the Training set and above 95% on the Validation set. This means that more than 99% of the 270000 train images were well predicted from epoch 10. Likewise, more than 95% of the 1800 validation slices were well predicted from epoch 10. Looking at the performance on the validation set, the model proto-DenseNet161 performed better than proto-VGG19 and proto-ResNet152 (see **Figure 3**).

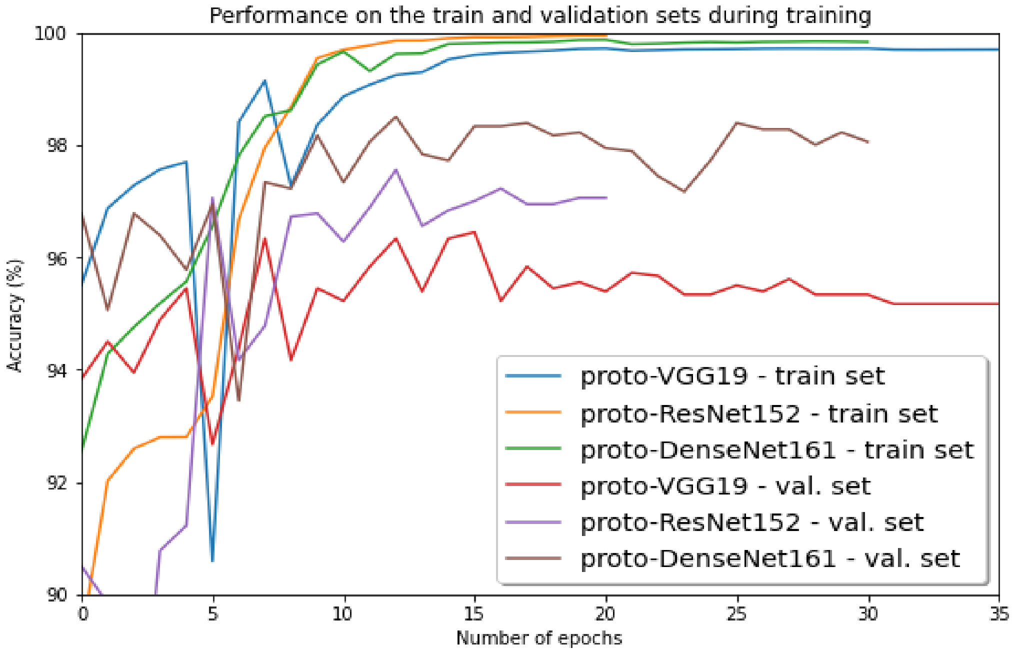


Figure 3. Evolution of model accuracies on the Training and Validation sets

3.3 Selection of the best model using ABIDE 1

We took the percentage of slices classified as corrupted (Class 1) as the probability that the whole scan is corrupted. For a given scan, if this percentage is >50%, then the predicted class of the scan was taken as Class 1. For a given scan, this threshold on the returned probability is used to produce a class prediction, because that is useful in the context of QC (pass/fail). However, there are some applications where an examination of the value of the probability itself might be warranted, since this may give more information about the quality of a scan or particular set of scans.

ProtoPNet Models: (CNN-base_epochs)	**Train - 60 scans**	**Validation - 12 scans**	**Test - 908 scans**

densenet161_10	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 69.82%
			ROC_AUC = 0.7751

densenet161_20	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 64.65%
			ROC_AUC = 0.7738

densenet161_30	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 62%
			ROC_AUC = 0.7578

resnet152_10	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 75.44%
			ROC_AUC = 0.8247

resnet152_20	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 68.72%
			ROC_AUC = 0.8107

vgg19_10	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 67.18%
			ROC_AUC = 0.8229

vgg19_20	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 70.04%
			ROC_AUC = 0.8494

vgg19_30	Accuracy = 100%
	ROC_AUC = 1
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 71.81%
			ROC_AUC = 0.8472

MRIQC_CLF	Accuracy = 96.67%
	ROC_AUC = 0.7667
		Accuracy = 100%
		ROC_AUC = 1
			Accuracy = 70.37%
			ROC_AUC = 0.7236

Table 2. Accuracy and ROC AUC scores for every ProtoPNet model on the Training, Validation, and Test sets. Last row: comparison with MRIQC performance.

Table 2 compares the classification accuracies for global quality of the Training, Validation, and Test sets, obtained for each of the models. The last row shows the accuracy for MRIQC scores launched on the same datasets. These results showed that the best model for the prediction of sMRI scan global quality is proto-ResNet152 trained on 10 epochs. This model has superior accuracy than MRIQC for the Training and test sets.

We should mention that to get the predictions from the MRIQC classifier, we did not set a particular threshold on probability values, we used the default parameters. Importantly, the MRIQC algorithm was trained using the ABIDE dataset, so its accuracy for the ABIDE dataset should be particularly good.

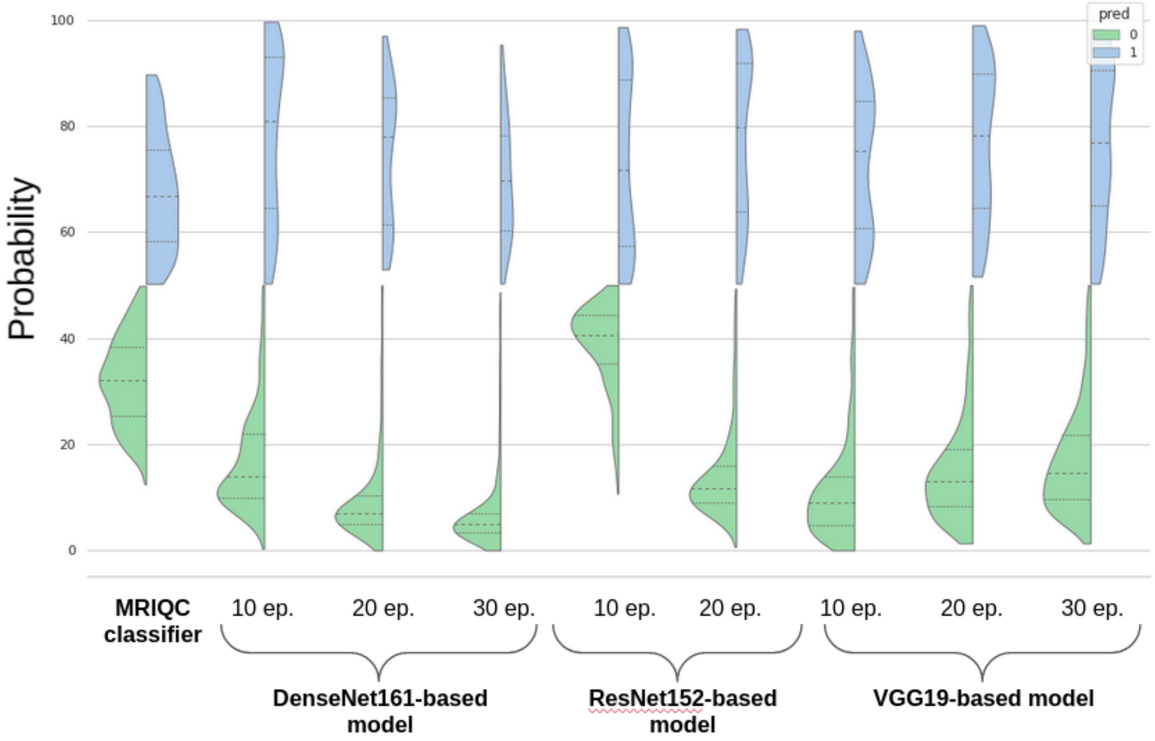


Figure 4. Comparison of the distribution of probabilities for the test set (908 scans), colored by predicted class: green for class 0 (good quality scans), blue for class 1 (medium/low quality scans).

In **Figure 4**, we can see that the distribution of predictions of "uncorrupted" (Class 0; green) scans looks gaussian for our models. In contrast, the distribution of predictions for "corrupted" (Class 1; blue) looks like a gaussian mixture. This distribution shape is expected since there are globally corrupted scans and locally corrupted scans, then the percentage of slices predicted to be corrupted will be different for the two types. In addition, there are different intensity levels for the artifacts as described by Backhausen et al. (2016) that might yield different levels of probability.

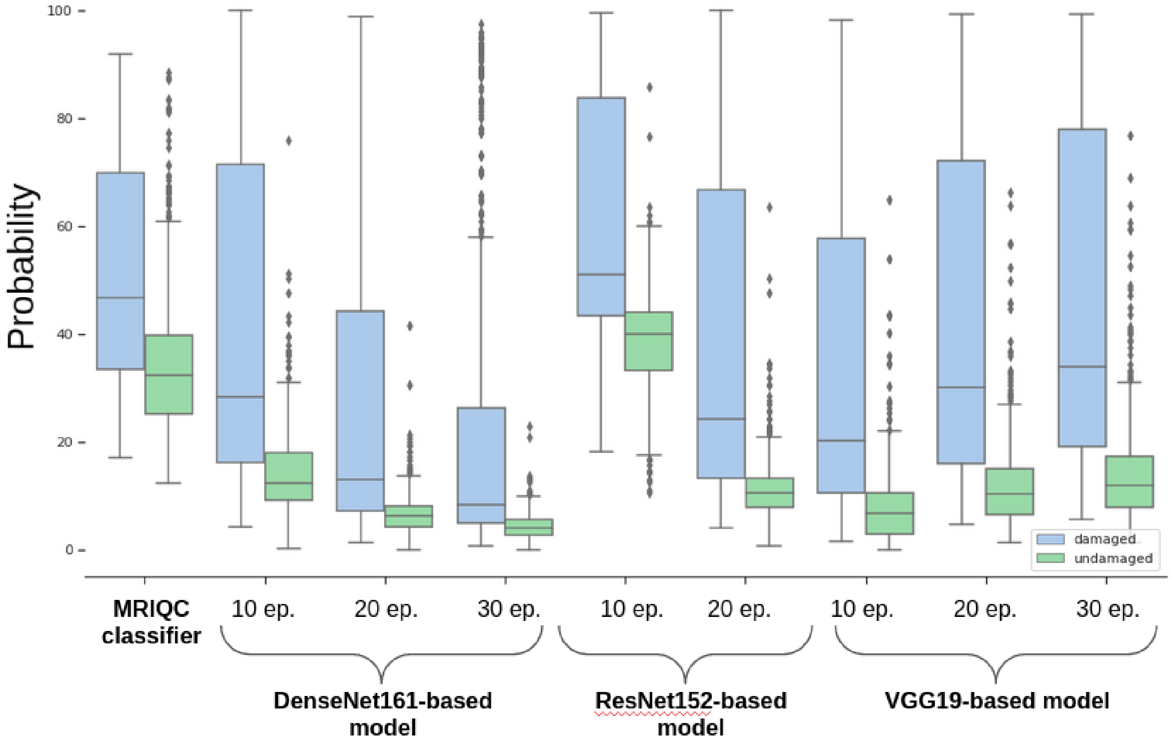


Figure 5. Boxplots show the predicted probabilities for truly good quality scans (green) and truly medium/low quality scans (blue) for all models and for MRIQC, using 980 scans from ABIDE 1.

Figure 5 shows that the probabilities of corrupted scans and those of uncorrupted scans are overlapping. The greater the overlap, the more False Positives and False Negatives there are. The overlap is greater for the MRIQC algorithm than for any of our models.

Test - 908 scans	Uncorrupted - Class 0 (528 scans)	corrupted - Class 1 (380 scans)
densenet161_10	Accuracy = 99.43%	Accuracy = 28.68%
densenet161_20	**Accuracy = 100%**	Accuracy = 15.53%
densenet161_30	**Accuracy = 100%**	Accuracy = 9.21%
resnet152_10[i]	Accuracy = 95.27%	**Accuracy = 47.89%**
resnet152_20[i]	Accuracy = 99.62%	Accuracy = 25.79%
vgg19_10[h]	Accuracy = 99.62%	Accuracy = 22.11%
vgg19_20[h]	Accuracy = 99.05%	Accuracy = 29.74%
vgg19_30[h]	Accuracy = 98.48%	Accuracy = 34.74%
MRIQC_CLF	Accuracy = 91.1%	Accuracy = 41.58%

Table 3. Accuracies for each class for every model and MRIQC on test sets

Table 3 compares the accuracy scores for prediction of each class separately. For Class 0 (good quality scans), all of our models have accuracy scores greater than 95%, while MRIQC has a lower score of 91.1%. For Class 1 (scans with artifacts), the scores are globally lower. The best score is achieved by the model proto-ResNet152 trained on 10 epochs (47.89%) followed by the MRIQC classifier (41.58%). These lower scores are explainable by the fact that, in the Test set, scans are less corrupted than in the Training set, and have different levels of intensity of artifacts. This might yield to probabilities between 0.4 and 0.5 for medium quality scans, meaning the class predicted is 0. This corresponds to the overlaps of probabilities shown in Figure 3. Moreover, for certain models, we might miss information because of the limited variety of prototypes randomly picked from the train set.

In addition, looking at the 2000 prototypes of each model, the set of prototypes of the model proto-ResNet152 - 10 epochs appeared to be the most diverse and relevant for the artifacts we annotated. Examples of such prototypes can be found in Section 9.1 of the Supplemental Information.

We deduced that proto-ResNet152 - 10 epochs was the best model among all the models tested in our experiment.

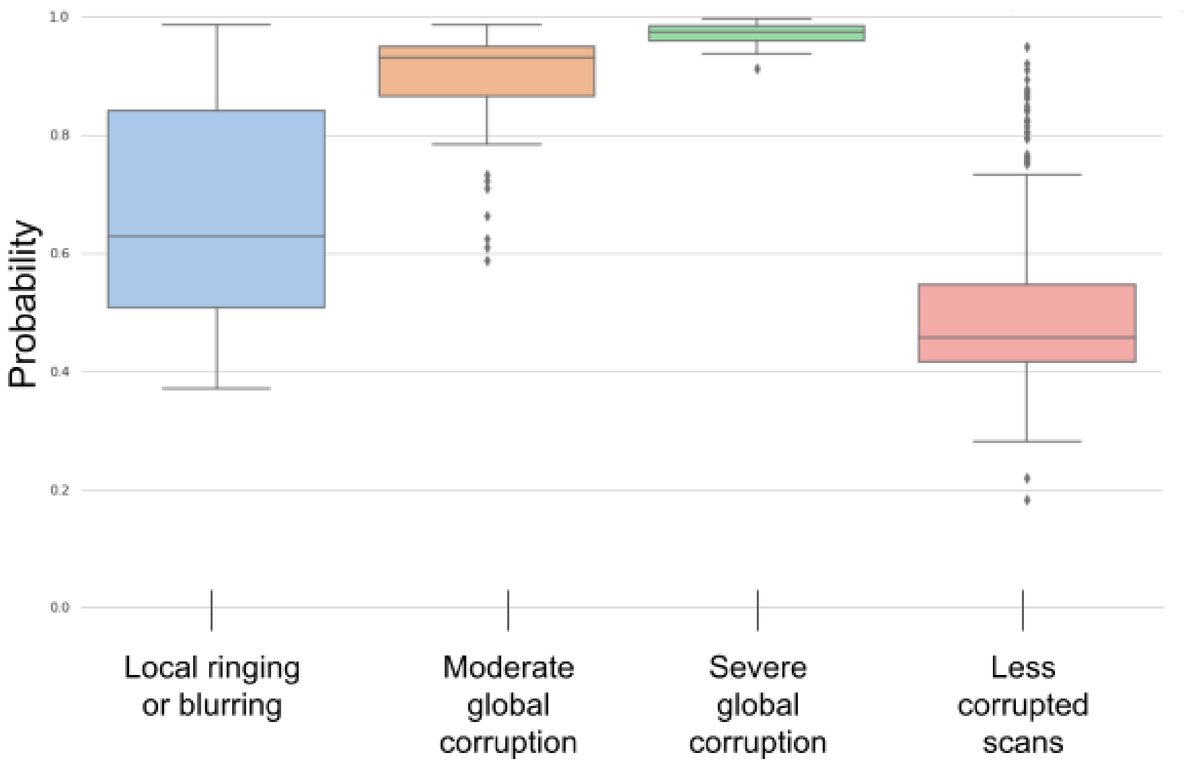


Figure 6. Comparison of probabilities from the model proto-ResNet152 trained on 10 epochs, on 416 scans with artifacts from ABIDE 1 (30 very low quality scans in train set, 6 very low quality scans in validation set, 380 globally or locally corrupted in test set). 51 scans have local ringing or blurring (blue), 60 are globally corrupted

but medium quality (orange), 36 are globally corrupted and very low quality (green), 269 are other less corrupted scans (red).

Figure 6 shows that different types of image artifact correspond to different predicted probabilities from our algorithm. The probabilities of globally corrupted scans are very close to 1, while the probabilities of scans with local ringing or blurring are spread between 0.5-0.8, and other, less corrupted scans have probabilities around 0.5.

We also evaluated the results on slices for the 66 scans from ABIDE 1 we annotated with local ringing and/or local blurring. We found that in the extremities, the algorithm tends to predict the slices as Class 1, even in the cases it should be Class 0. This means that slices near the edge of the field-of-view containing little brain tend to be identified as corrupted by the algorithm. This might explain why the global distribution of probabilities of the model proto-ResNet152-10ep is higher than the ones of other models (see **Figures 4** and **Figure 5**).

We also found an axis effect - while predictions for sagittal images were 89.3% accurate, accuracy for coronal images were 86.4% accurate, and for axial views, 78.8% accurate. We also found that it might be more difficult to automatically detect scans with only local areas of blurring because their global probabilities are comparable to the probabilities of uncorrupted scans, and local blurring was not always detected on slices. This might be due to fewer examples of this type of artifact in the training set and the set of prototypes.

Importantly, no site effect was observed (see **Table 4**), and there was no difference in the global distribution of probabilities between the three axes (sagittal, coronal, axial). These two points validate the approach of our model: using prototypes enables better generalisation of the algorithm to new data. Our findings show that our model is particularly efficient at detecting good quality scans and globally corrupted scans.

	good quality - 528 scans	**globally medium corrupted - 60 scans**	**local ringing or blurring - 51 scans**	**other less corrupted scans - 269 scans**
CALTECH	accuracy: 1.0 n scans: 34	na	na	accuracy: 0.0 n scans: 2
CMU	accuracy: 1.0 n scans: 24	na	na	accuracy: 0.3333 n scans: 3
KKI	accuracy: 1.0 n scans: 25	accuracy: 1.0 *n scans:* 3	na	accuracy: 0.5714 n scans: 14
LEUVEN_1	accuracy: 0.9259 n scans: 27	na	na	accuracy: 0.5 n scans: 2
LEUVEN_2	accuracy: 0.9565 n scans: 23	na	accuracy: 1.0 n scans: 1	accuracy: 0.2 n scans: 10
MAX_MUN	accuracy: 0.9286 n scans: 28	accuracy: 1.0 n scans: 2	accuracy: 1.0 n scans: 1	accuracy: 0.8 n scans: 10
NYU	accuracy: 0.9146 n scans: 82	accuracy: 1.0 n scans: 1	accuracy: 0.5882 n scans: 17	accuracy: 0.2714 n scans: 70

OHSU	accuracy: 0.9091 n scans: 22	accuracy: 1.0 n scans: 1	na	na
OLIN	accuracy: 0.75 n scans: 12	na	accuracy: 1.0 n scans: 2	accuracy: 0.4286 n scans: 7
PITT	accuracy: 0.9524 n scans: 21	na	accuracy: 1.0 n scans: 5	accuracy: 0.3913 n scans: 23
SBL	accuracy: 1.0 n scans: 26	na	na	accuracy: 0.0 n scans: 4
SDSU	accuracy: 0.8 n scans: 10	accuracy: 1.0 n scans: 10	na	accuracy: 0.8 n scans: 10
STANFORD	na	accuracy: 1.0 n scans: 10	accuracy: 0.8333 n scans: 12	accuracy: 0.6667 n scans: 6
TRINITY	accuracy: 1.0 n scans: 34	accuracy: 1.0 n scans: 3	accuracy: 1.0 n scans: 1	accuracy: 0.0 n scans: 7
UCLA_1	accuracy: 0.8958 n scans: 48	accuracy: 1.0 n scans: 6	accuracy: 0.6667 n scans: 3	accuracy: 0.8 n scans: 5
UCLA_2	accuracy: 1.0 n scans: 7	accuracy: 1.0 n scans: 3	accuracy: 1.0 n scans: 1	accuracy: 0.4286 n scans: 7
UM_1	accuracy: 1.0 n scans: 27	accuracy: 1.0 n scans: 7	accuracy: 0.8 n scans: 10	accuracy: 0.1471 n scans: 34
UM_2	accuracy: 1.0 n scans: 13	na	accuracy: 0.6667 n scans: 3	accuracy: 0.25 n scans: 12
USM	accuracy: 1.0 n scans: 60	na	na	accuracy: 0.25 n scans: 4
YALE	accuracy: 1.0 n scans: 5	accuracy: 1.0 n scans: 5	accuracy: 0.5 n scans: 4	accuracy: 0.1795 n scans: 39

Table 4. Predictions for each data collection site in the test set (908 scans) for the model proto-ResNet152 trained on 10 epochs.

3.4 Validation on ABCD (2141 scans) dataset

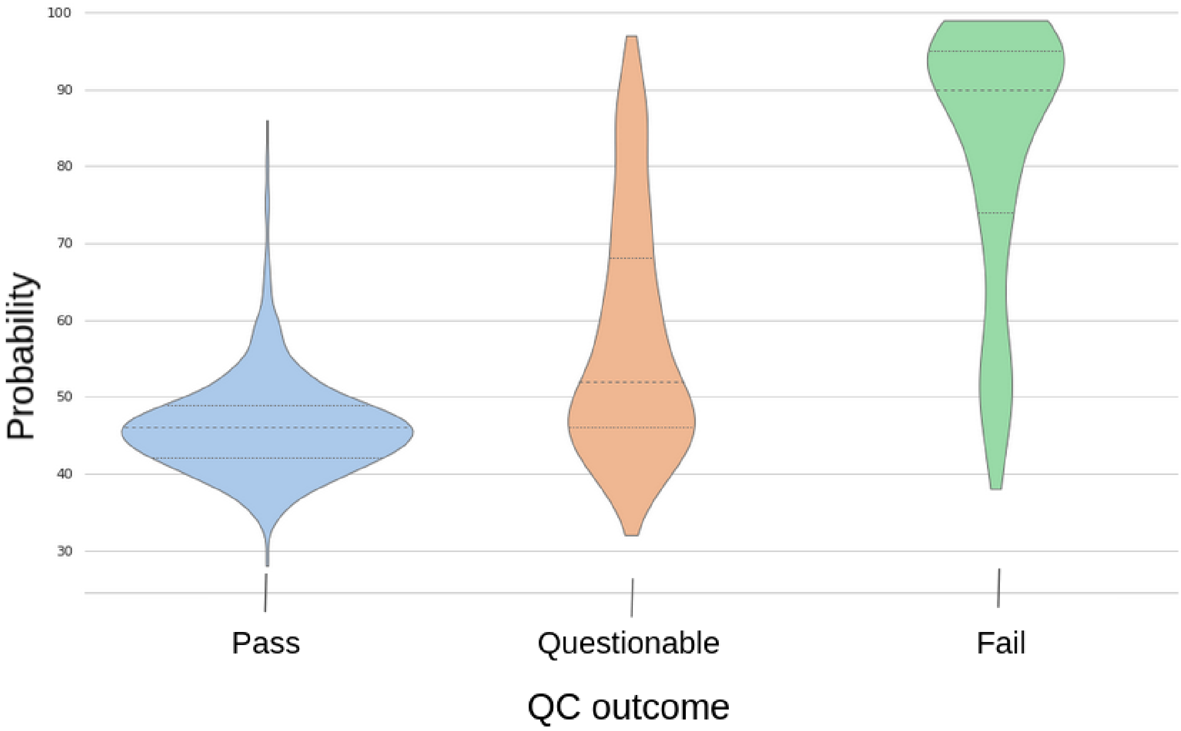


Figure 7. Comparing the distribution of probabilities between the true QC categories (pass, questionable, fail) for ABCD data (2141 scans), computed by proto-ResNet152 trained on 10 epochs.

ABCD (2141 scans)	pass	questionable	fail
proto-ResNet152 10 epochs	Accuracy = 82.4%	class 0: 255 class1: 304	Accuracy = 91.4%

MRIQC (on 410 scans only)	Accuracy = 90.4%	class 0: 43
		class 1: 7
			Accuracy = 76.1%

proto-ResNet152 10 epochs (removing scans with 0.5 < proba < 0.6)	Accuracy = 96.4%	class 0: 230
		class 1: 195
			Accuracy = 92.2%

Table 5. Accuracy of predictions for each of the manually determined QC categories (pass, questionable, fail) for ABCD data (2141 scans)

The ABCD dataset has been annotated with gold-standard manual QC judgments thanks to work groups facilitating data collection and quality control (Karcher and Barch, 2021). We tested our algorithm on 2141 of these manually QCed scans. **Figure 7** compares the distribution of probabilities between the true QC categories (pass, questionable, fail) for these 2141 ABCD scans, computed by proto-ResNet152 trained on 10 epochs. It shows that the scores are globally distinct between pass and fail categories, while scores of the category questionable are in-between, which is expected.

Table 5 shows that our algorithm showed better accuracy in predicting the category "fail" (accuracy of 91.4% versus 76.1% for MRIQC). Conversely, MRIQC initially performed better than proto-ResNet152 when predicting the category "pass" (accuracy of 90.4% versus 82.4%). Upon closer inspection, we found that our algorithm predicted 311 scans with probabilities between 0.5 and 0.6, mostly belonging to the category pass. When these scans are removed and only scans with probabilities lower than 0.5 or greater than 0.6 are retained, accuracy was 96.4% for the pass category. We therefore recommend a second verification (manual or with MRIQC) for scans with "borderline" probabilities from our algorithm, between 0.5 and 0.6.

Among the questionable scans, 108 scans were annotated "use this one", and our algorithm predicted 66% of these scans in class 0 (good quality scans).

Supplementary validation of the algorithm using ABIDE 2 (799 scans) and on ADHD-200 (751 scans) is included in Section 9.3 (Supplemental Information)**.**

3.5 BIDS Docker app

We developed a BIDS-app (Gorgolewski et al., 2017) to share our model with the neuroimaging community. It is available on the open-source platforms GitHub and DockerHub. It is ready-to-use by following the instructions on: https://github.com/garciaml/BrainQCNet.

The optimal version is the one that is compatible with GPU compatible with CUDA. We found that the average time to process a 3D sMRI scan was about 1 minute on a laptop with one GPU Nvidia GEFORCE GTX 1060. Processing a scan with the MRIQC algorithm took about 45 minutes on the same machine.

There is also a version available to work on CPU machines. We found that the average time to process a scan was about 20 minutes on a laptop with Intel Core I7 processor.

4. Discussion

In this age of “big data”, manual quality control of T1-weighted MRI scans is a time-consuming task requiring substantial experience and training. Our goal was to further advance the automatic detection of artifacts in structural brain MRI T1-weighted scans. We trained a Deep Learning algorithm - ProtoPNet - with several different architectures - VGG19, ResNet152, DenseNet161 - to classify good and poor quality scans. Our results indicate that the best model was able to detect poor quality scans very well, whatever the architecture of the convolutional layer architecture. It also predicted high quality scans very well. For scans with more localized rather than global artifacts, the specific slices containing artifacts are also well detected by our models.

Across architectures, ProtoPNet with ResNet152 CNN architecture trained on 10 epochs showed the best performance. On the first testing set (908 scans from ABIDE 1 (Di Martino et al., 2014)), this model showed better performance in predicting the global class of a scan than the reference tool, MRIQC (accuracy for high quality scans: 95.27% vs 91.1% for MRIQC; accuracy for medium and low quality scans: 47.89% vs 41.58% for MRIQC). We also showed that the overlap between the distributions of probabilities (percentage of slices classified as corrupted/Class 1) for good quality scans and the distributions for scans with artifacts is much reduced with our model, which demonstrates that, in the training dataset, our model better discriminates between scans with artifacts and scans without.

On the second testing set (2141 scans from ABCD; Volkow et al., 2018; Karcher and Barch, 2021), our proto-ResNet152 model showed excellent accuracy for medium and low quality scans: 91.4% vs 76.1% for MRIQC). MRIQC tended to have more False Negatives than our model in the sub-dataset tested. For high-quality scans, our model showed very good prediction accuracy (82.4%), but this was lower than that found for MRIQC (90.4%). When we examined this more closely, we found that the mid-range of probabilities [0.5;0.6] predicted by our model contained a mixture of good quality scans and moderately corrupted scans with more localized artifacts. If this range is excluded, our model exhibits excellent accuracy for both high- and low-quality classes (accuracy for high quality scans: 96.4% ; accuracy for low quality scans: 92.2%). Accordingly, we suggest that the specific threshold may need to be adjusted according to the needs of your study. Here, we set it at 0.5, such that scans with probabilities >0.5 were predicted low quality, and scans with probabilities <0.5 were predicted as high quality. If a researcher had a very generous sample and wanted to retain only the very best quality scans, the threshold could conservatively be set at 0.5 - this would have the disadvantage of removing some relatively good quality scans but the advantage of ruling out 91.4% of low quality scans. If, on the other hand, a researcher had a smaller sample and less stringent quality requirements (e.g., is not performing analyses of brain volume cortical thickness), a more liberal threshold of 0.6 could be set. This would mean that some scans with delimited areas of poor quality would be included in the study, but would offer the advantage that no good quality scans would be unduly eliminated. A third possibility is for researchers to retain all scans that have a global probability lower than 0.5, and to manually evaluate or run MRIQC on scans that have a global probability between 0.5 and 0.6 to separate the good from moderately corrupted scans. Conveniently, it is possible to manage this threshold on our app, as we explain in the documentation (https://github.com/garciaml/BrainQCNet), which also explains how to use the app even if your data is not BIDS-structured.

In addition to increasing the accuracy of QC, our study demonstrates that Deep Learning is a promising method for increasing the speed of scan quality evaluation while reducing the computational resources required. To generate a global prediction for a single 3D scan on a GPU machine, our model currently takes 1 minute to process one scan (vs. 40-45 minutes with MRIQC). On a CPU machine, our model is slower but still relatively fast (20 minutes to process one scan). Although the intermediates created by MRIQC processing may be used in further analyses of the data, this processing is arguably wasteful of resources in the case of categorically poor quality scans and large datasets. In obviating long processing time, our method is potentially more sustainable. In order to further save resources and encourage sustainable practices, we have shared the global scores predicted by our best model for the scans we used from ABIDE 1 and 2 (Di Martino et al., 2014; Di Martino et al., 2017), ADHD200 (Bellec et al., 2017) and ABCD (Volkow et al., 2018; Karcher and Barch, 2021). The scores are available through our github repository: https://github.com/garciaml/BrainQCNet.

Another potential benefit of our model is its higher level of interpretability. The local detection of corruption might help to identify specific regions that have a greater susceptibility to artifacts. This may, for example, highlight a scanner quality issue that can be addressed, a brain area that is particularly vulnerable to motion, or, in the case of a clinical group, it may suggest the need for specific interventions to avoid data loss. We have made it easy to inspect regions exhibiting local artifacts using our app. This involves reorienting your image to the canonical space RAS+ and using the parameter “n_areas” of our app to inspect the probabilities that artifacts are present in different areas of the image. More details on how to proceed can be found in the documentation (https://github.com/garciaml/BrainQCNet).

Future work includes the improvement of this algorithm by running more experiments with other CNN-bases like ResNet34 or DenseNet121 and examining the effects of prototype selection. In addition, we plan to increase the training set as well as the variety of artifacts in the set of prototypes. Investigating whether our approach could be applied to other MRI modalities is another important future direction. Quality Control of fMRI is a huge challenge that is exacerbated by the advent of Big Data. Future work will examine whether our approach can be adapted for data with a temporal dimension so that it could be applied to fMRI data in a framewise manner to enable faster and automated data quality control. Finally, to our knowledge, our BIDS-app is the first app that applies Deep Learning to neuroimaging and is built to be used on CUDA GPU machines. By sharing our code, we are providing the community with a new BIDS-app template for Deep Learning applications, facilitating the sharing of Deep Learning models in the community and helping to maximize reproducibility and collaboration.

5. Conclusions

In this work, we introduced a novel Deep Learning approach for the automatic evaluation of the quality of brain structural MRI scans. Our method is scalable to big datasets by taking advantage of new technologies like GPU machines with high-computing capacity. Our results highlighted the reliability and the relevance of our Deep Learning model in assessing the global quality of 3D brain T1-weighted scans, being stable across differences in acquisition protocols. It also showed satisfying detection of artifacts at the local level. Paths to improve our model include trying to combine CNN architectures, or manually selecting the prototypes for the model. This approach could be further adapted to functional MRI,and to other types of scans and organs.

Our model is already freely available for the global assessment of the quality of brain structural MRI scans by the community via the app BrainQCNet (https://github.com/garciaml/BrainQCNet). Since all our code is open-source, the app can be used as a template for future applications of Deep Learning in Neuroimaging.

6. Acknowledgements and Funding

This research was supported by the Hypercube Institute (Paris, France), a Trinity College Dublin Postgraduate Research Studentship (1252) awarded by the School of Medicine, and an Irish Research Council Postgraduate Scholarship

7. Disclosure of competing interests

None.

8. References

Alfaro-Almagro, F., et al., 2018. *Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank.* NeuroImage 166, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Backhausen, L.L., et al., 2016. *Quality Control of Structural MRI Images Applied Using FreeSurfer—A Hands-On Workflow to Rate Motion Artifacts*. Front. Neurosci. 10. https://doi.org/10.3389/fnins.2016.00558

Bellec, P., et al., 2017. *The Neuro Bureau ADHD-200 Preprocessed repository.* NeuroImage 144, 275–286. https://doi.org/10.1016/j.neuroimage.2016.06.034

Chen, C.,et al., 2019. *This Looks Like That: Deep Learning for Interpretable Image Recognition.* arXiv:1806.10574 [cs, stat].

Clark, D.B., et al., 2018. *Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience*. Developmental Cognitive Neuroscience 32, 143–154. https://doi.org/10.1016/j.dcn.2017.06.005

Deng, J., et al., 2009. *ImageNet: A large-scale hierarchical image database*, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), IEEE, Miami, FL, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Di Martino, A., et al., 2017. *Enhancing studies of the connectome in autism using the autism brain imaging data exchange II*. Sci Data 4, 170010. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., et al., 2014. *The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.* Mol Psychiatry 19, 659–667. https://doi.org/10.1038/mp.2013.78

Esteban, O., et al., K.J., 2017. *MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites*. PLoS ONE 12, e0184661. https://doi.org/10.1371/journal.pone.0184661

Gilmore, A., Buser, N., Hanson, J.L., 2019. *Variations in Structural MRI Quality Significantly Impact Commonly-Used Measures of Brain Anatomy* (preprint). Neuroscience. https://doi.org/10.1101/581876

Glasser, M.F., et al., 2016. *The Human Connectome Project's neuroimaging approach.* Nat Neurosci 19, 1175–1187. https://doi.org/10.1038/nn.4361

Gorgolewski, K.J., et al., 2017. *BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods*. PLoS Comput Biol 13, e1005209. https://doi.org/10.1371/journal.pcbi.1005209

Gorgolewski, K.J., et al., 2016. *The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments*. Sci Data 3, 160044. https://doi.org/10.1038/sdata.2016.44

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

He, K., et al., 2015. *Deep Residual Learning for Image Recognition.* arXiv:1512.03385 [cs].

Huang, G., et al., 2018. *Densely Connected Convolutional Networks.* arXiv:1608.06993 [cs].

Karcher, N.R., Barch, D.M., 2021. *The ABCD study: understanding the development of risk for mental and physical health outcomes*. Neuropsychopharmacol. 46, 131–142. https://doi.org/10.1038/s41386-020-0736-6

Keshavan, A., Yeatman, J.D., Rokem, A., 2019. *Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging*. Front. Neuroinform. 13, 29. https://doi.org/10.3389/fninf.2019.00029

Khanh, T.L.B., et al., 2020. *Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging*. Applied Sciences 10, 5729. https://doi.org/10.3390/app10175729

LeCun, Y., et al., 1999. *Object Recognition with Gradient-Based Learning,* in: Shape, Contour and Grouping in Computer Vision, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–345. https://doi.org/10.1007/3-540-46805-6_19

Marcus, D.S., et al., 2013. *Human Connectome Project informatics: Quality control, database services, and data visualization*. NeuroImage 80, 202–219. https://doi.org/10.1016/j.neuroimage.2013.05.077

Nordahl, C.W., et al., 2016. *Methods for acquiring MRI data in children with autism spectrum disorder and intellectual impairment without the use of sedation*. J Neurodevelop Disord 8, 20. https://doi.org/10.1186/s11689-016-9154-9

Ranjbarzadeh, R., et al., 2021. *Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images.* Sci Rep 11, 10930. https://doi.org/10.1038/s41598-021-90428-8

Rauch, S.L., 2005. *Neuroimaging and Attention-Deficit/Hyperactivity Disorder in the 21st Century: What to Consider and How to Proceed.* Biological Psychiatry 57, 1261–1262. https://doi.org/10.1016/j.biopsych.2005.02.014

Reuter, M., et al., 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

Simonyan, K., Zisserman, A., 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 [cs].

Sudlow, C., et al., 2015. *UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.* PLoS Med 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sujit, S.J., et al., 2019. *Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks.* J. Magn. Reson. Imaging 50, 1260–1267. https://doi.org/10.1002/jmri.26693

Volkow, N.D., et al., 2018. *The conception of the ABCD study: From substance use to a broad NIH collaboration*. Developmental Cognitive Neuroscience 32, 4–7. https://doi.org/10.1016/j.dcn.2017.10.002

Whelan, C.D., et al., 2018. *Structural brain abnormalities in the common epilepsies assessed in a worldwide ENIGMA study.* Brain 141, 391–408. https://doi.org/10.1093/brain/awx341

White, T., et al., 2018. *Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction.* Hum. Brain Mapp. 39, 1218–1231. https://doi.org/10.1002/hbm.23911

Zarrar, S., et al., 2015. *The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data*. Front. Neurosci. 9. https://doi.org/10.3389/conf.fnins.2015.91.00047

Zhang, N., et al., 2014. *Part-based R-CNNs for Fine-grained Category Detection*. arXiv:1407.3867 [cs].

Zheng, H., Fu, J., Mei, T., Luo, J., 2017. *Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition*, in: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, pp. 5219–5227. https://doi.org/10.1109/ICCV.2017.557

Zhou, B., et al., 2016. *Learning Deep Features for Discriminative Localization*, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 2921–2929. https://doi.org/10.1109/CVPR.2016.319

9. Supplemental Information

9.1 Examples of activation maps

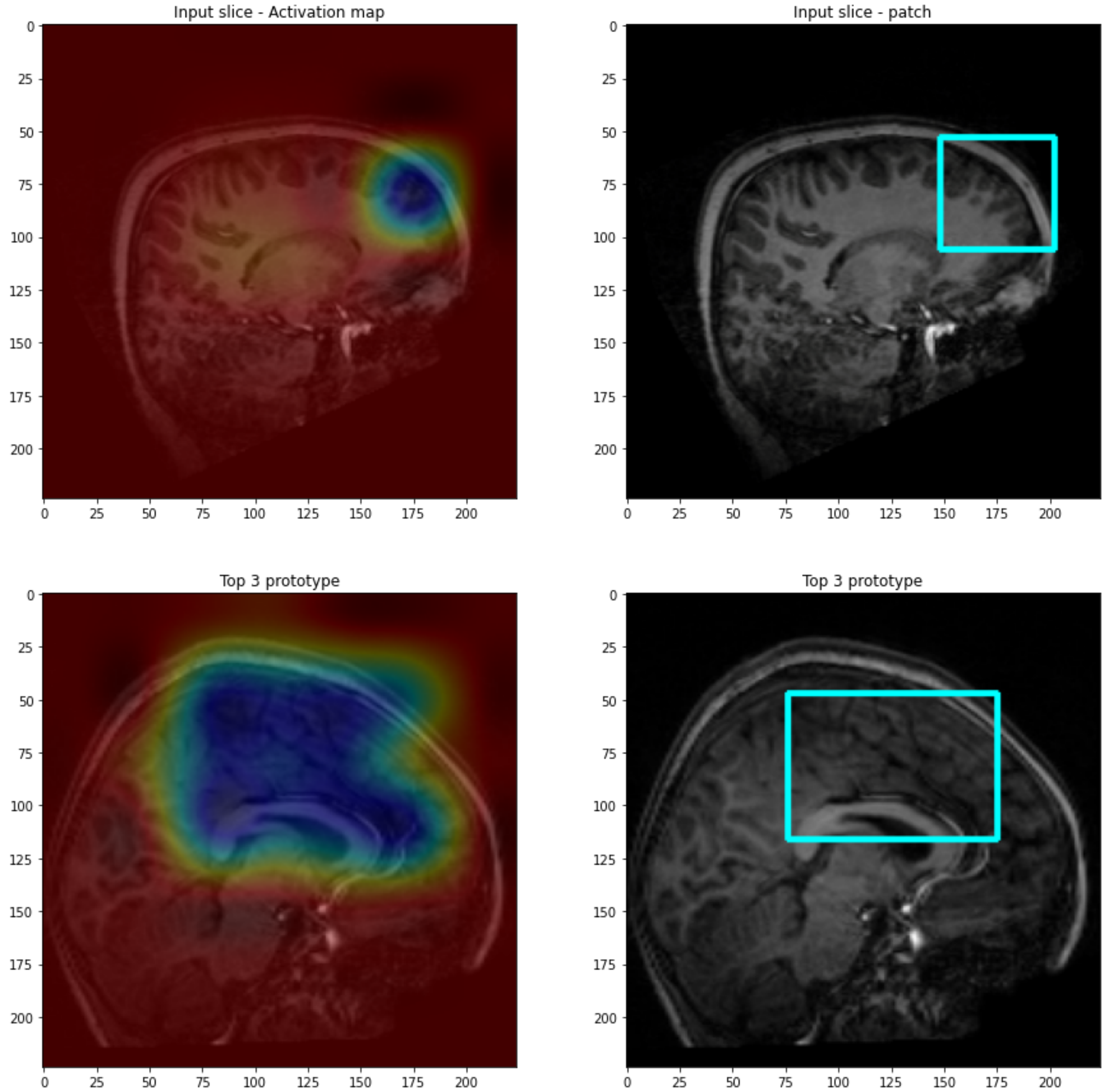


Figure 9. Examples of meaningful artifact map and prototype: the upper panel shows the input slice, the lower panel shows the top-3 prototype for the model proto-RESNET152 trained on 10 epochs.

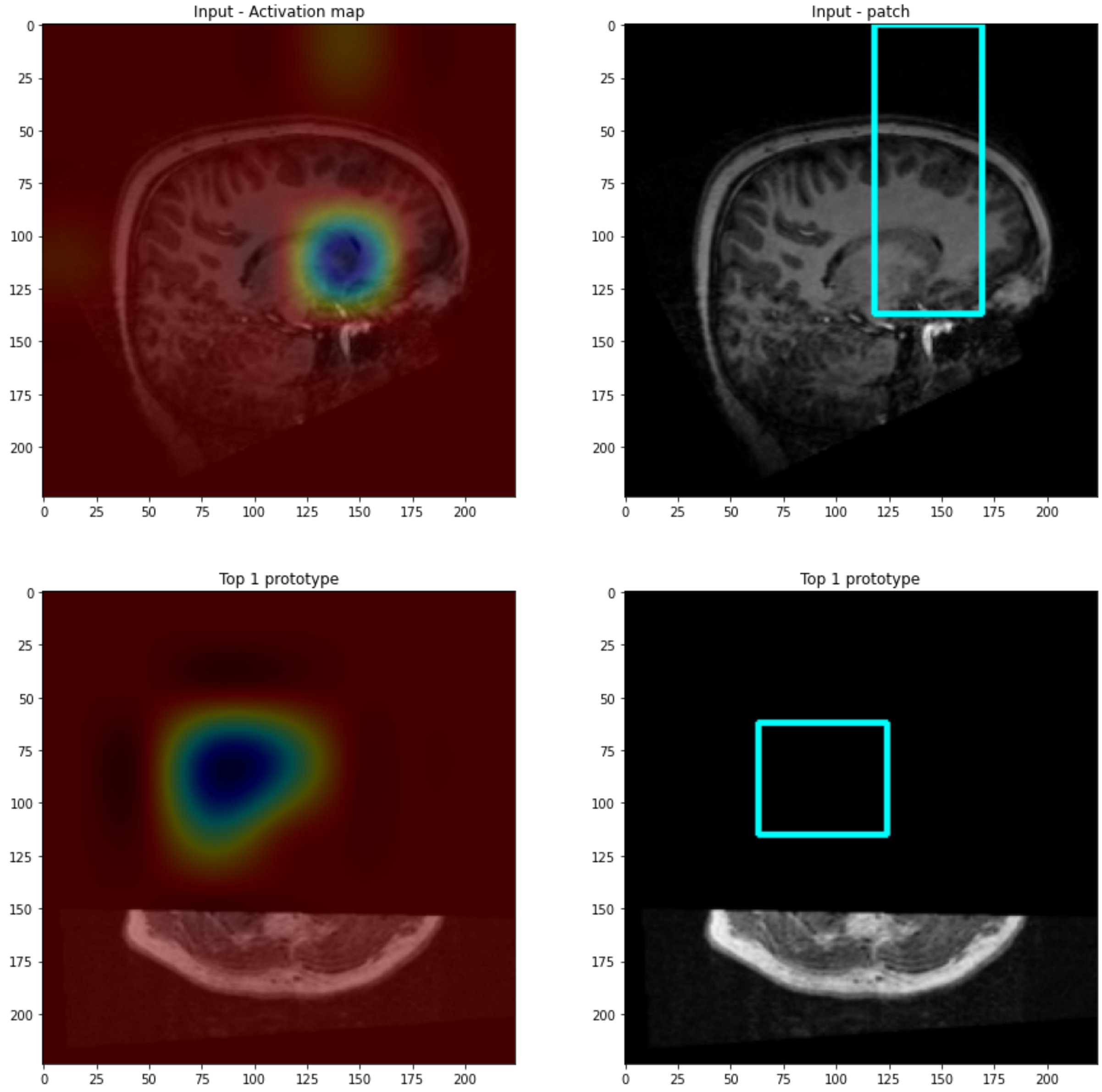
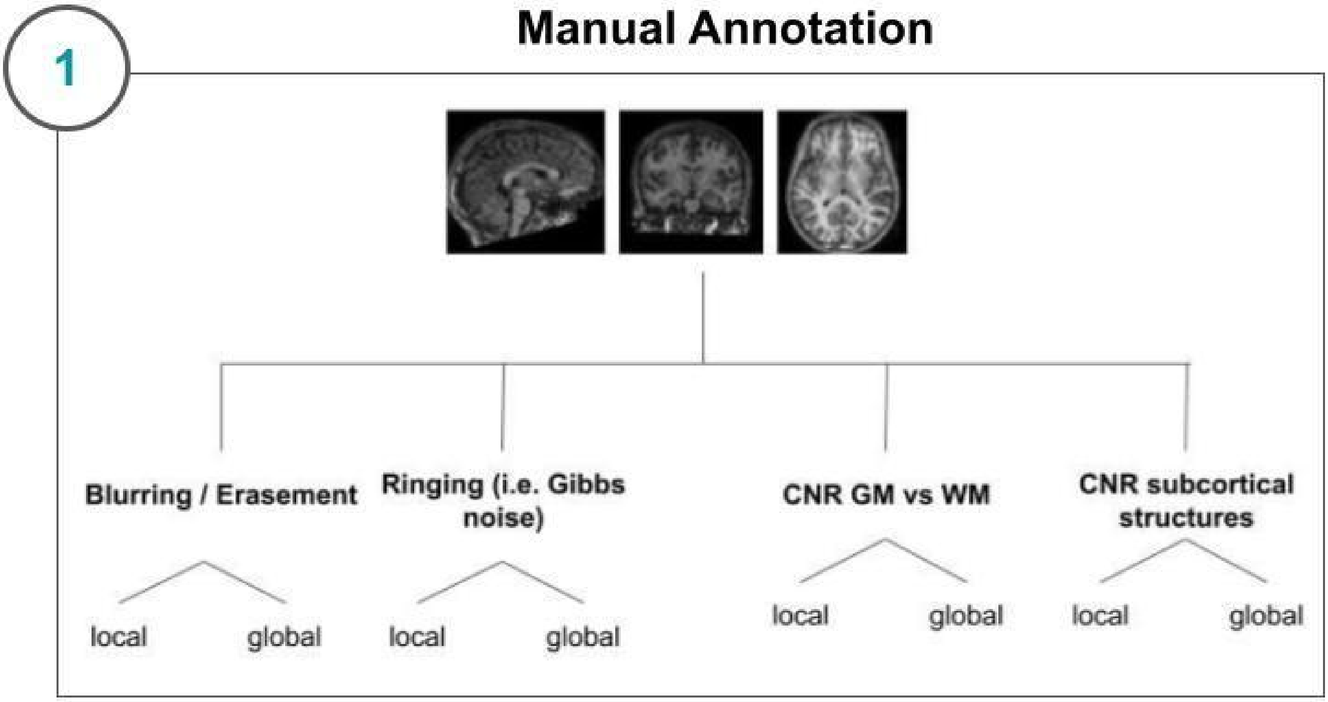
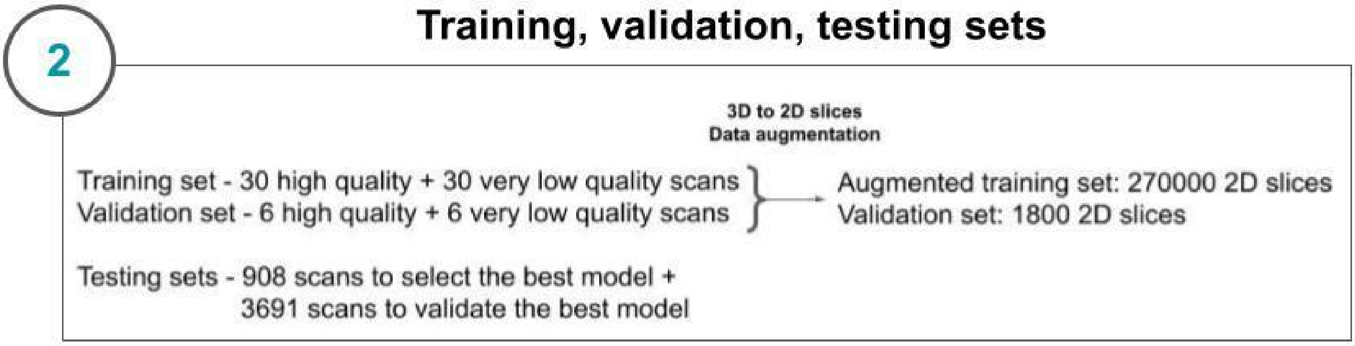


Figure 10. Examples of non-meaningful artifact map and prototype: the upper panel shows the input slice, the lower panel shows the top-1 prototype for the model proto-VGG19 trained on 30 epochs.

9.2 Pipeline steps





Training set - 30 high quality + 30 very low quality scans
Validation set - 6 high quality + 6 very low quality scans

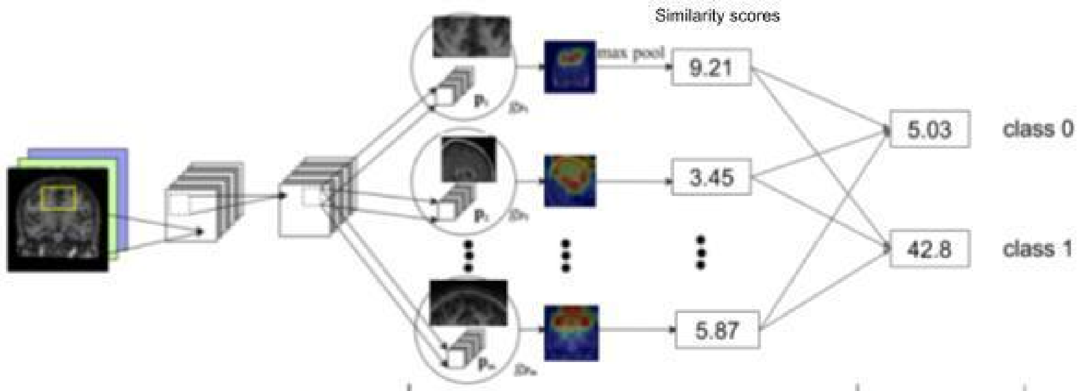
Augmented training set: 270000 2D slices
Validation set: 1800 2D slices

Testing sets - 908 scans to select the best model +
3691 scans to validate the best model

3

Building Models - Training (augmented training set and validation set)

Iterative optimization process;
models taken at 10, 20 and 30 epochs.



Convolutional layers

Prototypes layer

Fully connected layer

Output logits

3 architectures tried:

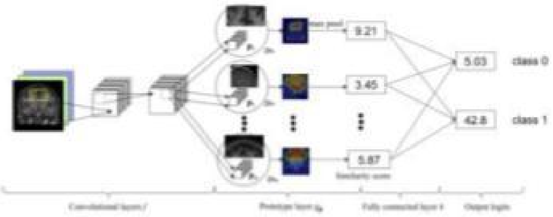
- VGG19
- ResNet152
- DenseNet161

4 & 5

Best model selection and validation (with testing sets)

Inference of predictions

Slice 1



Slice 2

...

Slice N

3D scan = Slices 1 to N

Global probability (percentage of slices predicted 1)

Global prediction

ROC AUC

Accuracy

9.3 Supplementary validation using ABIDE 2 (799 scans) and ADHD-200 (751 scans).

To further validate our tool using 799 scans from ABIDE 2 dataset, we ran the MRIQC classifier on this dataset and treated the results as ground truth. We obtained an accuracy score of 75.5% and a ROC AUC score of 0.72. Taking the MRIQC classifier results as ground truth introduces a bias since in their paper they showed that they had an accuracy score around 75% on dataset including ABIDE. However, our result shows that our algorithm tends to predict the quality of scans well.

The dataset ADHD200 provided manual annotations for 751 scans. We ran our algorithm on this dataset, and obtained an accuracy score of 79.2% and a ROC AUC score of 0.76. These results also show that our algorithm predicts the quality of scans reliably.