Computers in Biology and Medicine 151 (2022) 106211

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Application of a convolutional neural network to the quality control of MRI defacing

Daniel J. Delbarre [a,b,*], Luis Santos [a,b], Habib Ganjgahi [c], Neil Horner [a], Aaron McCoy [a], Henrik Westerberg [a], Dieter A. Häring [d], Thomas E. Nichols [e,f], Ann-Marie Mallon [a,b]

[a] MRC Harwell Institute, Harwell Campus, Oxfordshire, OX11 0RD, United Kingdom
[b] The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, United Kingdom
[c] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, United Kingdom
[d] Novartis Pharma AG, Basel, Switzerland
[e] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford, OX3 7LF, United Kingdom
[f] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom

ARTICLE INFO

Dataset link: 10.5281/zenodo.6638765

Keywords:
Magnetic resonance imaging
Defacing
Quality control
Neural network
Anonymisation

ABSTRACT

Large-scale neuroimaging datasets present unique challenges for automated processing pipelines. Motivated by a large clinical trials dataset with over 235,000 MRI scans, we consider the challenge of defacing — anonymisation to remove identifying facial features. The defacing process must undergo quality control (QC) checks to ensure that the facial features have been removed and that the brain tissue is left intact. Visual QC checks are time-consuming and can cause delays in preparing data. We have developed a convolutional neural network (CNN) that can assist with the QC of the application of MRI defacing; our CNN is able to distinguish between scans that are correctly defaced and can classify defacing failures into three sub-types to facilitate parameter tuning during remedial re-defacing. Since integrating the CNN into our anonymisation pipeline, over 75,000 scans have been processed. Strict thresholds have been applied so that ambiguous classifications are referred for visual QC checks, however all scans still undergo an efficient verification check before being marked as passed. After applying the thresholds, our network is 92% accurate and can classify nearly half of the scans without the need for protracted manual checks. Our model can generalise across MRI modalities and has comparable performance when tested on an independent dataset. Even with the introduction of the verification checks, incorporation of the CNN has reduced the time spent undertaking QC checks by 42% during initial defacing, and by 35% overall. With the help of the CNN, we have been able to successfully deface 96% of the scans in the project whilst maintaining high QC standards. In a similarly sized new project, we would expect the model to reduce the time spent on manual QC checks by 125 h. Our approach is applicable to other projects with the potential to greatly improve the efficiency of imaging anonymisation pipelines.

1. Introduction

A collaboration between Novartis and the University of Oxford's Big Data Institute (BDI) has been established to improve drug development and healthcare through the application of artificial intelligence and advanced analytics. Large, multidimensional datasets – consisting of clinical, imaging and omics data – are integrated and analysed to improve patient prognoses and identify early predictors of disease. A dedicated research informatics framework has been developed that allows this data to be captured, anonymised, explored and integrated into databases [1]. Currently, the collaboration focuses on two main therapeutic areas: Multiple Sclerosis (MS), and autoimmune diseases treated with the interleukin (IL)-17 A antibody secukinumab (IL17). Data from over 50,000 patients is available across both projects, and the MS project will utilise both clinical and brain magnetic resonance imaging (MRI) data. The MRI scans are a key data source for the MS project, with over 235,000 scans from over 12,000 unique subjects. Many of the subjects have longitudinal MRI data over several years and corresponding clinical information, enabling the progression of disease to be studied [2].

Before the MRI data can be used for research purposes it is necessary to homogenise the data to a single format, and to anonymise all identifying patient data, including metadata and any facial features

* Corresponding author at: The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, United Kingdom.
E-mail address: ddelbarre@turing.ac.uk (D.J. Delbarre).

https://doi.org/10.1016/j.compbiomed.2022.106211
Received 7 January 2022; Received in revised form 26 August 2022; Accepted 9 October 2022
Available online 18 October 2022

that are present in the image data. While MRI metadata (which is stored in DICOM tags in raw MRI data) can be readily anonymised, through deletion or modification of selected tags, the anonymisation of the image data itself is more complex. Two commonly used approaches are skull stripping and defacing. Skull stripping involves removing all non-brain tissue from a scan, and can be implemented using a number of methods [e.g. 3–5]. However, skull-stripping methods often require considerable fine-tuning when applied to large datasets containing scans of variable image quality [6], and additionally, skull-stripped images can sometimes be unsuitable when similar algorithms are required for processing images downstream of de-identification work [7]. Defacing techniques retain non-brain tissue and can be implemented through shearing off parts of the face [e.g. 8], blurring the outer surface of the face [9] and selectively removing areas of the face that contain identifiable facial features [e.g. 10–12].

Quality control (QC) processes are commonly employed when working with MRI data to ensure that the data is suitable for downstream analysis, and that artefacts in the data will not introduce bias into analyses [13]. Furthermore, the quality of defacing may also need to be QC checked [e.g. 14] to ensure that not only are any identifying facial features correctly removed from the scan, but also to ensure that the brain anatomy has not been damaged. The choice of defacing method is important as the performance of the software can vary between scans, particularly when the data has been acquired at different sites, and with different acquisition parameters [15]. Additionally, the choice of software can impact the results of downstream analyses [16,17]. If high standards of QC checks are not employed, then there is a possibility that patients could be identified through photographic visual comparisons [18], facial recognition software [17,19,20], and the facial reconstruction of inadequately defaced MRI scans [17,21]. Many defacing methods have been developed on high resolution, research quality scans. In this collaboration, which utilises MRI scans from global clinical trials, the scans are typically lower resolution, were captured from a large number of sites, and due to the longitudinal nature of the data, some scans were acquired over 15 years ago. Therefore, due to the potentially high levels of variation in the quality of the data, there is greater potential for variation in the successful application of defacing to these scans. As a consequence of this, thorough QC checks are necessary to ensure that data is correctly anonymised.

While visual QC checks are commonly employed to assess the quality of MRI scans, when undertaking projects containing tens of thousands, or in the case of this project hundreds of thousands of scans, these manual checks become impractical. The time-consuming nature of the checks can cause considerable delays between receiving data and having the data research-ready. Furthermore, those undertaking the QC may become fatigued and more likely to make errors. Citizen scientists can be used to assist with visually QC checking MRI data [e.g. 22], but this is not suitable when anonymisation checks are being undertaken. Automated methods have been developed to assist with the QC of MRI data, utilising methods including univariate classifiers [23], support vector machines (SVMs) [24], random forests [25], neural networks [26], or a combination of classifiers [11,27]. Automated QC methods may not perform adequately on all data — meaning that it may still be necessary to conduct some manual QC checks. However, allowing automated methods to handle the clear-cut cases allows for manual QC to be focused on data that is more difficult to classify. Additionally, automated methods must be generalisable so that they can process heterogeneous data. Despite the prevalence of automated models used to QC MRI data, these are typically focused on QC checking image quality or analysis output. The QC of MRI defacing normally requires visual human checks to ensure that the facial features have been fully removed. With the exception of one recent binary classifier [28], there is a complete lack of automated methods to assist with defacing QC checks. However, as deep learning approaches are now widely applied to other MRI QC tasks, there are substantial opportunities to also improve the efficiency of defacing QC checks by applying similar approaches.

In this work, we developed a convolutional neural network (CNN) to assist with the QC checking of defaced MRI images from a large dataset containing images from numerous sources and of variable quality. Our network was developed using 24,000 pre-classified renders of defaced scans. Following development of the CNN, we evaluate model performance to select strict probability thresholds that convey high levels of accuracy, allowing for reliable automatic classification and for manual QC checks to be targeted towards problematic scans. We also describe the integration of the CNN into a pre-existing image anonymisation pipeline, complementing our existing manual QC processes, and including the adoption of time-efficient verification checks to protect patient anonymity. Furthermore, we evaluate the implications of implementing the CNN with regard to the time that is saved in comparison to fully manual QC checks, and discuss the potential for the adoption of machine learning approaches to improve the efficiency of MRI anonymisation pipelines.

2. Material and methods

2.1. MRI data

MRI data was available from 24 Novartis clinical studies, with the entire dataset containing over 235,000 MRI scans from over 12,000 subjects. A detailed overview of the dataset has been reported by Dahlke et al. [2]. The majority of scan sessions contained T1-weighted (T1w; with and without gadolinium contrast enhancement), T2-weighted (T2w) and proton density (PD) modalities. Some subjects also had T1-weighted (T1w) 3D acquisitions, diffusion-weighted (DWI), fluid-attenuated inversion recovery (FLAIR), and magnetisation transfer (MT) scans. MRI scans were of variable quality as they were captured from over 1000 unique MRI scanners, with some scans captured over 15 years prior to the start of the collaboration. The majority of scans had a 256×256 or 192×256 acquisition matrix, with 46 or 60 slices. Prior to anonymisation, facial features are readily discernible in these scans, with the exception of the DWI scans in which facial features are not visible. It was a requirement for our project that MRI scans had to be de-identified before they could be made available to analysts.

2.2. MRI defacing pipeline prior to CNN development

MRI data was initially transferred to a dedicated, secure anonymisation environment for defacing and the anonymisation of any remaining confidential patient data. The anonymisation environment was only accessible to a very small team of dedicated scientists not otherwise involved in the research. A bespoke pipeline (Fig. 1) was built to handle the processing of this data into a consistent format, ready for downstream research. Data was initially converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format using the DICOM conversion software *HeuDiConv* [29], and structured using the Brain Imaging Data Structure standard (BIDS) [30] - a standard for brain MRI datasets, that is widely used within the neuroimaging research community. Anonymised metadata was preserved in a JSON file that accompanied each NIfTI file.

During the conversion process, a select number of DICOM tags are extracted and added to the JSON files by the converter; these tags provide metadata that may be required for downstream analysis. The conversion software has a preset list of tags to extract, which includes tags recommended for inclusion by the BIDS specification. Some tags that were included by *HeuDiConv* were completely removed from all JSON files; this included some free-text fields (e.g. 'Image Comments') and details that could identify institutions. Unique values for JSON tags were reviewed to ensure that no un-anonymised data had inadvertently been included in the JSON files.

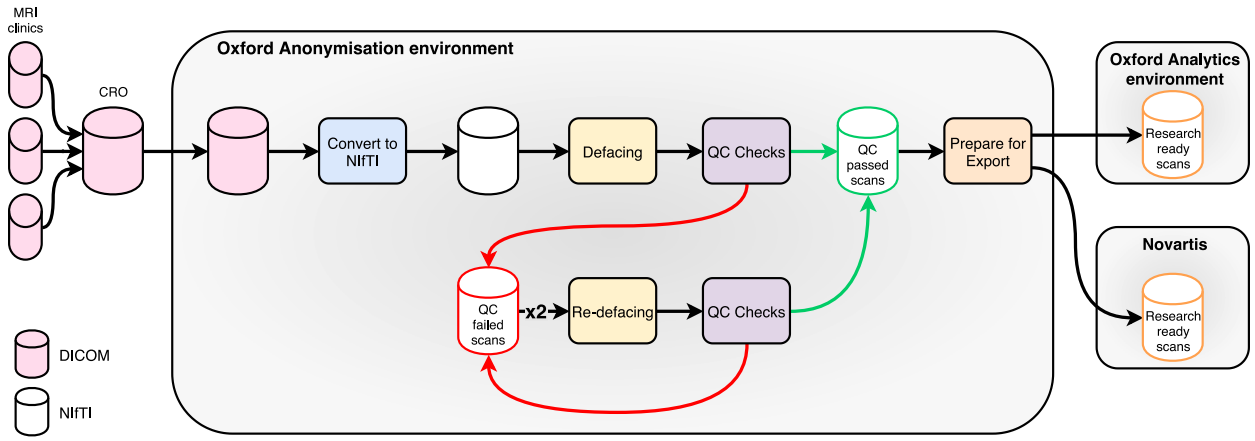

Fig. 1. Flowchart showing the anonymisation pipeline for MRI data.

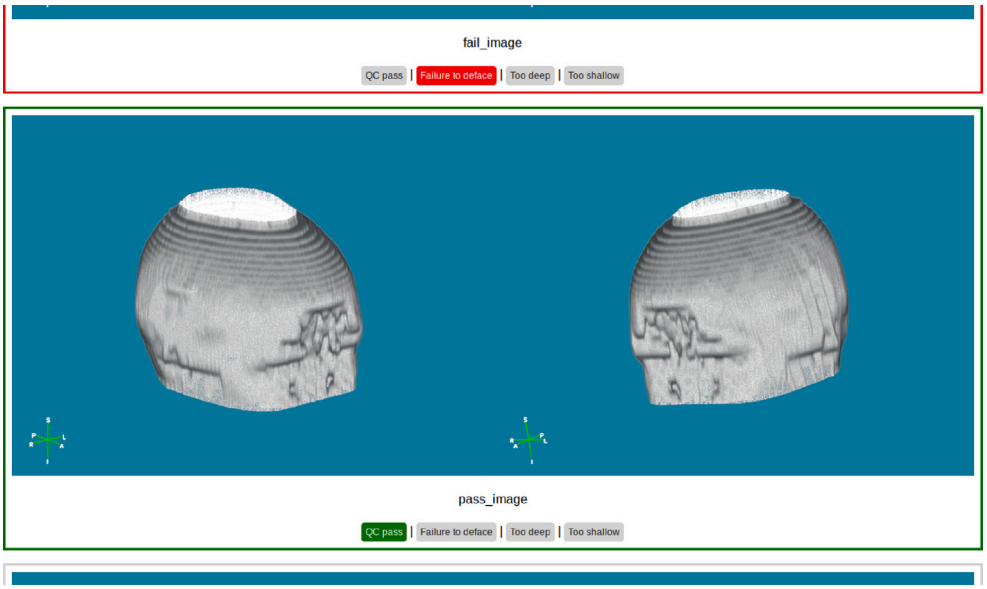

Fig. 2. Usage of the html page when undertaking manual QC checks. Note that the renders in this figure are generated from a publicly-available sample scan from [31].

After conversion to NIfTI format, each scan was defaced. During this process, identifying facial features, specifically the ears, eyes, nose and mouth, are removed from each scan. Defacing was implemented using the *fsl_deface* software [11], which is known to perform well at preventing re-identification and minimise interference with downstream analysis methods when compared to other defacing software [17]. Unlike many of the other available defacing methods, *fsl_deface* removes the ears as well as other facial features. Following being defaced, all scans were QC checked to ensure that the defacing had been applied correctly. Scans were classified as one of four categories: 'pass' (scan defaced correctly), 'deep' (defacing went too deep and encroached on the brain), 'shallow' (defacing did not go deep enough and facial features were still visible), 'failure' (broad category for complete registration failures and scans containing unfixable errors; defacing failures that do not fit into the 'deep' and 'shallow' categories). Prior to the development of the neural network, all QC checks were carried out manually, using a HTML page generated by a Python script. Each of the scans were visualised as two PNG images of 3D renders of the scan (left and right oblique views) which were generated using the *fsl_gen_3D* software [11,32]. Using the HTML page, each scan was classified by clicking the button that corresponds to one of the above categories (Fig. 2). After a batch of QC had been completed, CSV files recording the classifications were saved.

After the initial QC checks were completed, remedial defacing was undertaken on scans which did not pass, a process that we called re-defacing. During re-defacing, custom parameters were selected by considering the type of scan and the previous QC classification(s). All the custom parameter combinations were configurable with the options available in *fsl_deface*. When defacing was classified as too deep or too shallow, the defacing mask was nudged anteriorly or posteriorly respectively during the re-defacing. The application of bias correction and/or reducing the fractional intensity was used when a defacing attempt had been classified in the 'failure' category. The re-defacing was also applied in an additive manner. For example, if a scan had been classified as 'shallow' during initial QC checks, and was still classified as 'shallow' after the first round of re-defacing, then the mask would be nudged even more posteriorly during the second round of re-defacing. Following each round of the re-defacing, scans were once again QC checked using the same protocol as with the initial defacing. Up to two rounds of re-defacing were undertaken to ensure that as many scans as possible passed the defacing stage. On average, after initial defacing 69% of scans had passed QC checks, but 96% of scans were successfully defaced after two rounds of re-defacing. All anonymised NIfTI scans, and accompanying JSON metadata files, that passed the QC checks were then prepared for export to a separate, analysis-focused environment at the BDI and also back to Novartis, ready to be included

in downstream research and analysis. Any scans that did not pass the QC checks did not leave the anonymisation environment and were not included in any downstream analysis.

2.3. CNN: Data selection and pre-processing

As over 100,000 scans had been put through our anonymisation pipeline and manually QC checked prior to the development of the neural network, a large quantity of labelled scan data was available for use in developing the CNN. However, not all of this data could be used in network development for two reasons. Firstly, as the majority of prior manual QC classifications were for scans classified as 'pass' (75%) and to a lesser extent 'shallow' (15%), there was considerably less 'deep' (6%) and 'failure' (4%) classification data, so in order to keep the proportion of data from each class balanced (at least initially; see below) a smaller subset of the available scans were used. Secondly, as most subjects usually have multiple scans (usually ≥8, but up to 60) and some of these scans had been re-defaced and therefore QC checked multiple times, it was preferable not to use all available scans to reduce the chance of the CNN overfitting to the anatomy of subjects appearing in the dataset multiple times.

Initially, 16,000 scans were used to develop the model, with 4000 scans per class, which were split 60:20:20 between training, validation and test sets. Due to the initial models performing better on the 'deep' and 'failure' classes, the proportion of 'shallow' and 'pass' scans in the dataset was increased to 8000 scans per class, giving 24,000 scans in total. Scans from each class were randomly selected from those available to represent the variability found in the entire dataset. This included scans of varying image quality, which were selected from multiple clinical studies and the available modalities.

A single 2D image was used as the input for the CNN. Each image was a horizontal concatenation of the two renders used in the manual QC checks, which were cropped to remove parts of the background that did not contain parts of the anatomy in any scans. The rationale for using the 2D renders over the 3D scans was as follows: the average NIfTI scan in this project contains over 3 million voxels, most of which is not data that is applicable to assessing the quality of the defacing and would be computationally expensive to process in its entirety. Training needed to be computationally efficient as GPUs were not available within the anonymisation environment, and it was not possible to move the data externally for this purpose. In addition, while image registration could be used to extract the head/face region (and downsample the 3D data), defacing failures are often the result of poor image registration. Therefore, by using 2D images of the renders, visual information that is pertinent to the effectiveness of the defacing is retained in a more compact format, and the style of the renders is consistent regardless of image quality. The renders were input into the CNN as RGB data for the following reasons. Firstly, the renders are generated with a coloured background for the manual QC checks; not having to convert these renders to greyscale streamlines the integration of the CNN into the defacing pipeline. Secondly, it is possible that the CNN could use the coloured background to detect edges of the render, including cases where the defacing mask has been applied incorrectly leaving large holes in the scan. Thirdly, the impact of using three channels was minimal on performance, with only a moderate increase in training time (~10%).

2.4. CNN: Network development

Utilisation of computational resources for the development of the neural network within the anonymisation environment had to be balanced with the requirements for concurrently running the existing defacing pipeline; the conversion and defacing steps are particularly computationally demanding. Additionally, only having CPUs available for training placed some limitations on network design. All CNN development was undertaken using the *keras* library [33] for R [34]. A transfer learning approach was initially used, where the convolutional base of a pre-trained VGG-16 model [35] was left frozen, and the densely-connected classifier was unfrozen. As the performance of the transfer learning model was sub-optimal (see Results), a simplified version of the Alexnet CNN [36] was also used. Like the VGG model, Alexnet has also previously won the high-profile ImageNet classification challenge [37], which contains over 1.2 million training images split into 1000 classes. As we had a much smaller number of classes, and less training images, we used a simplified version of the model to account for the reduction in the complexity of the task, to aid generalisability, and for compatibility with computational resources. A number of initial models were run to help define the upper and lower bounds of the parameter space that would be used for tuning the model, and this also ensured that the architecture was sufficient for classifying the data. The images were input as 165 × 270 × 3 (height × width × depth) tensors with values scaled between 0–1. After the initial runs showed that the models were overfitting, image augmentation (zoom, horizontal flipping) and dropout layers were used in all subsequent model development; augmentation was not applied to validation or test data. The training data was also shuffled. The parameter space included the number of convolutional layers (3, 4, 5), the number of nodes per layer (starting at 32 or 64 in the first convolutional layer), learning rate (0.00001, 0.0001, 0.001), and dropout rate (0.2, 0.5). A grid search approach was used to tune the parameters; models were generated in an iterative manner from the combinations in the parameter space. The final parameter set was selected by comparing the validation performance of all of the possible models from the parameter space and selecting the combination with the best performance.

The final model (Fig. 3) contained four convolutional layers, each using a rectified linear unit (ReLU) activation function and a kernel size of 3 × 3. Each convolutional layer was followed by a max pooling layer to downsample the data. After flattening, two dropout and three densely connected layers were interspersed. Both dropout layers had a dropout rate of 0.2. The first two dense layers used ReLU activation, with the final dense layer using softmax activation with an output size of four — one for each of the four classes. The network was optimised using a root mean square optimiser (RMS) with the learning rate set to 0.00008; 0.0001 was selected during tuning, but this was manually reduced to further optimise performance. Categorical cross-entropy was used as the loss function. The network was trained for 100 epochs, with 576 steps and a batch size of 25, so that each epoch contained all 14,400 training images. During each epoch the model was evaluated on the validation set, with 192 steps and a batch size of 25 used. A callback was used during training so that at the end of each epoch the model would be saved if its validation performance was superior to the previously saved best performing model; minimised validation loss was used as the criterion. After the training had been completed, the best performing model was run on the test dataset to evaluate its performance and calculate metrics.

The assignment of classifications was visualised using Gradient-weighted Class Activation Mapping (Grad-CAM) [38]. This approach allowed for the activation of a class for an input image to be visualised, which shows what features of the MRI renders the CNN is using to assign classifications. The Grad-CAM method was implemented in R using the approach of Chollet et al. [39], where a semi-transparent heatmap of the predicted class activations from the final convolutional layer is superimposed on top of the input image. For illustrative purposes, a publicly-available sample T2w MRI dataset [31] was used for the images included here; the parameters of the defacing script were modified to produce an example of each of the four classes.

Multi-class receiver operating characteristic (ROC) curves and the corresponding area under the curves (AUC) were calculated using the method of Hand and Till [40], as implemented in the *pROC* R package [41].

As well as evaluating the performance of the model at predicting the four classes, we also evaluated the binary performance of the model at

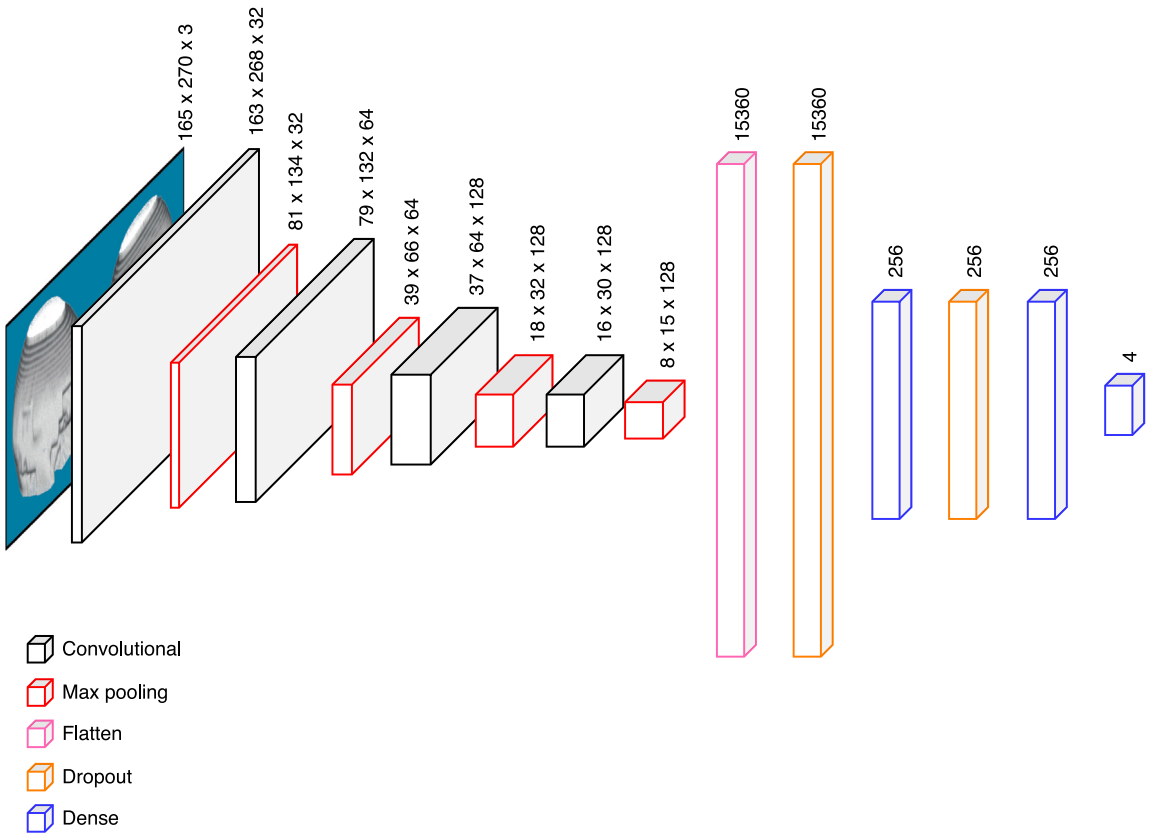

Fig. 3. Network diagram of the CNN.

identifying 'pass' scans versus all defacing errors combined. To do this, the 'deep', 'shallow' and 'failure' classifications were combined into one 'non-pass' class. This allowed us to investigate the CNN's ability to identify correctly defaced scans compared to those with any sort of defacing error. We used the 'non-pass' category as the positive class when evaluating the binary performance.

2.5. Probability threshold selection

Before the network was incorporated into the defacing pipeline, cut-off probability thresholds were selected so that only classifications with high confidence would be accepted. The adoption of strict cut-off thresholds was necessary to ensure that inadequately defaced scans were not marked as 'pass', and that incorrectly defaced scans were allocated to the correct class to allow for the appropriate parameters to be applied during remedial re-defacing work. Performance across classes was evaluated using the test dataset to select thresholds that will deliver acceptable performance upon implementation into the defacing pipeline. Cut-off probability thresholds were evaluated using metrics including sensitivity, specificity, and volume of data that would be classified at different thresholds (between 0.25–1.00 at 0.05 increments) for each class. Classifications were assigned based on the class with the highest probability. During the implementation phase (see below), classifications which did not surpass the threshold were discarded and these scans were QC checked manually instead. Therefore, the metrics used to assess the impact of applying different thresholds only include scans with accepted classifications. For example, if at threshold × classifications for 600 (out of 1000) scans reached the threshold, the sensitivity and specificity would be calculated for only those 600 scans that would be classified, and not the entire set of 1000 as we know that the remaining 400 scans will be manually QC checked.

2.6. Model evaluation on a dissimilar dataset

The publicly available IXI MRI dataset [42] was used to evaluate the performance of the CNN on a dissimilar dataset to the one on which it was developed on. The IXI dataset contains MRI scans from 582 subjects across five modalities and have not been defaced. The IXI scans have been used in other defacing studies [e.g. 21]. We used 200 randomly selected scans from each of the T1w, T2w and PD modalities to evaluate the model's performance. The scans were defaced using the *fsl_deface* software, initially using the default parameters. As there was a poor distribution between the classes, the IXI scans were also defaced using a refined parameter set (fractional intensity = 0.4, mask nudged posteriorly 5 mm). This provided a better distribution of scans between the four classes, allowing for a more representative comparison between the performance of the CNN on the IXI data to that of the study data. The 2D images of the renders were generated in the same way as described above. Each scan was manually QC checked prior to running the CNN. The model's performance was evaluated by comparing the manual QC classifications with those generated by the CNN (both pre and post application of the thresholds).

2.7. Integrating the CNN into the defacing pipeline

The CNN was integrated into the defacing pipeline (Fig. 4), and was used at the initial defacing and the two re-defacing stages. Following the defacing of all the scans in a study, a concatenated image of the two renders was generated to match the format of the images that were used in training the network. Classifications for the three non-'pass' categories were accepted when they exceeded the cut-off probability threshold for a class. Pass classifications that surpassed the 'pass' cut-off threshold were preliminarily accepted, pending a visual verification check before being fully accepted. The visual verification check is a quicker version of the full manual QC check, where the images are

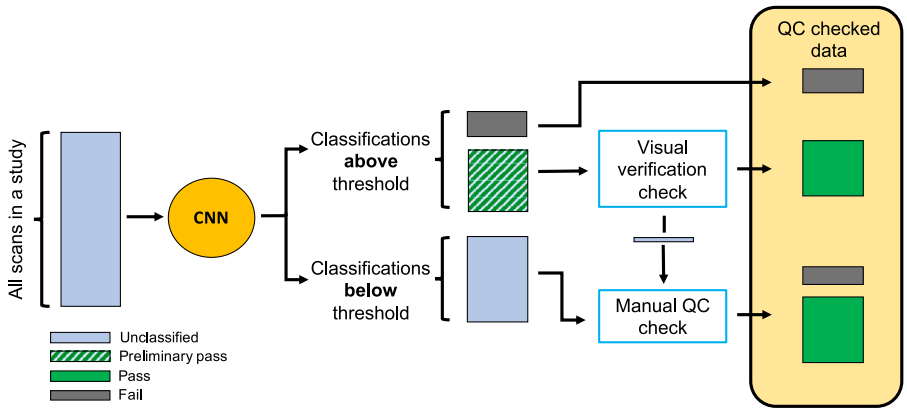

Fig. 4. Flowchart showing how the CNN has been integrated into the anonymisation pipeline. This process takes places within the 'QC checks' part of the anonymisation pipeline (shown in Fig. 1).

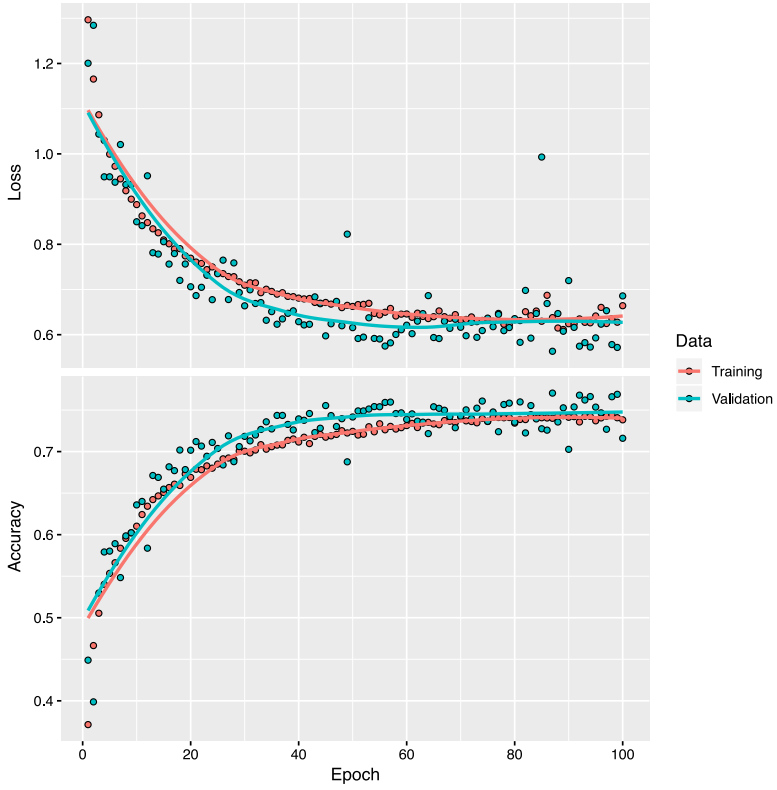

Fig. 5. Plots showing loss and accuracy during training the CNN. The loss and accuracy are shown for both the training and validation data at the end of each training epoch.

displayed in a gallery view that can be swiftly browsed and only those images where the 'pass' classification is not agreed with are flagged. Any scans that were flagged during the verification check, and the remainder of the scans that did not reach any of the cut-off thresholds, were then QC checked using the original manual method.

Integration of the CNN into the pipeline required no additional computational resources. Running the CNN is easily initiated by running a wrapper script. The wrapper script firstly calls an R script that prepares the data, runs the model, restructures the data based on the assigned classifications, and produces CSV files recording the assigned classifications and probabilities. The wrapper then runs a Python script to generate a HTML page for the visual verification checks. After the verification checks have been completed, another wrapper script is used to amend the classification for any scans that were flagged, and then generate HTML pages for any scans that still required the manual QC checks.

After the CNN was incorporated into the pipeline, its impact was assessed relative to the baseline manual QC process that was used prior to development of the CNN. The change in the efficiency of the QC process from using the CNN was determined by calculating the time that would be saved for a hypothetical study containing 10,000 scans. The performance of the CNN (e.g. the proportions of scans classified in each of the four classes, the proportion of CNN assigned passes accepted during verification checks) was ascertained by evaluating its performance following integration into the pipeline. The time taken to perform the manual and verification checks was determined by calculating the average time it takes to check a single scan in a batch of 250 scans.

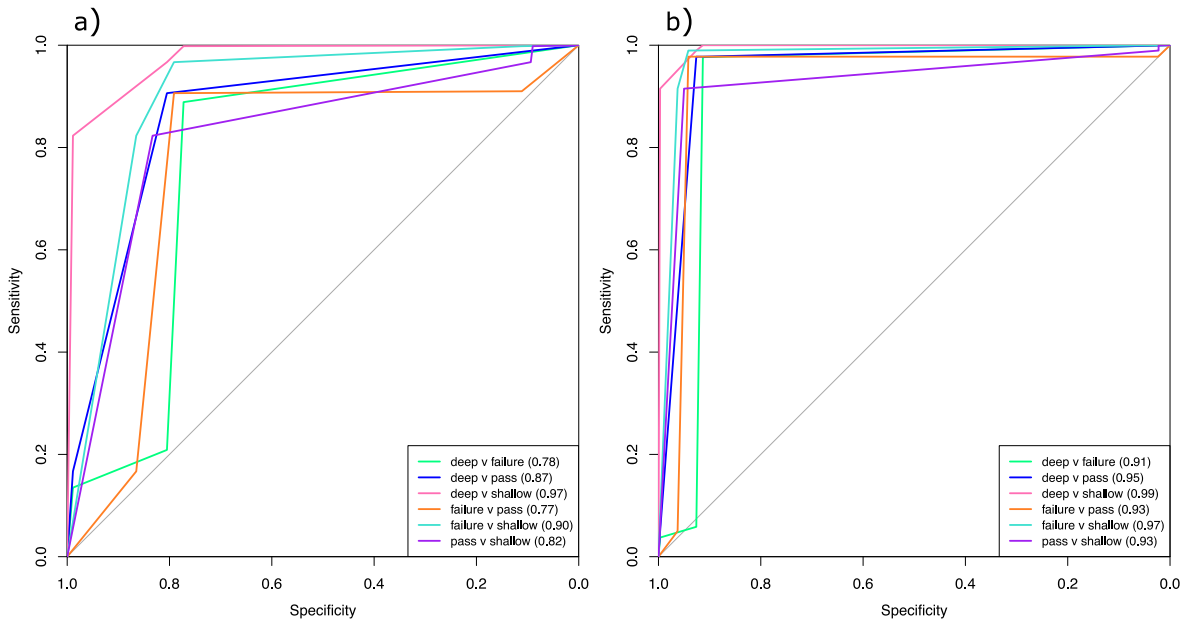

Fig. 6. Multi-class ROC curves when (a) all classifications are accepted, and (b) when the thresholds are applied. Each curve represents one pair-wise comparison. The AUC is shown in the legend for each comparison.

3. Results

3.1. Network performance

During model training, both the training and validation loss steadily decreased over 60 epochs, after which the loss decreased more incrementally, before plateauing after around 90 training epochs (Fig. 5). Overall, the training and validation loss were very similar throughout training, although the training loss was slightly higher due to the addition of dropout regularisation to the model. The training loss was lowest at epoch 89, where it reached 0.61, while the validation loss went down to 0.56 at epoch 87. Training and validation accuracy steadily increased until around epoch 60 before slowing and plateauing. Training accuracy reached 0.74 at epoch 76, and validation accuracy was highest at epoch 89, reaching 0.77. The loss and accuracy were very consistent during training, with only three epochs producing noticeable outliers in the training metrics. As the best performing network on the validation set, the model from epoch 89 was retained. For comparison, during model tuning other iterations of the model had validation accuracy ranging between 0.64 to 0.75. The best transfer learning model had a validation accuracy of 0.71.

When the test set was evaluated using the CNN (Table 1), accuracy of 0.76 with a 0.56 loss was achieved; this is similar to the values obtained during training on the validation set (0.77 accuracy and 0.56 loss). The sensitivity varied between the classes, ranging from 0.66 for the 'failure' class up to 0.82 for the 'shallow' class. The specificity is relatively high for all four classes (≥0.87). The model performs very well when distinguishing between 'deep' and 'shallow' classes (AUC = 0.97; Fig. 6a), and performs reasonably well distinguishing between 'failure' and 'shallow' classes (AUC = 0.90), and between 'deep' and 'pass' classifications (AUC = 0.87). The CNN performs poorest when distinguishing between 'failure' and 'pass' data (AUC = 0.77) .

When the CNN's binary performance at identifying 'pass' versus 'non-pass' ('deep', 'failure' and 'shallow' combined) classifications is evaluated (Table 2), the CNN's accuracy is greater (0.82) than when it is used for multi-class classification. The sensitivity is 0.87 and the specificity is 0.73.

When the test set is broken down by modality (Table 3), the performance of the CNN is generally consistent although there are some noticeable differences. Accuracy ranges between 0.71 in the T2w scans up to 0.80 in the T1w gadolinium scans. In most cases the sensitivity is >0.70, but it is very low for the T1w 'failure' class (0.50) and the T2w 'pass' class (0.54), but was highest for the T2w 'shallow' class (0.88). The specificity is generally high across the modalities (>0.85), especially in the 'deep' and 'failure' classes across all of the modalities (>0.90) but is lowest with the 'shallow' class for the T2w scans.

The features that the CNN uses to assign the classes can be visualised using the Grad-CAM heatmaps (Fig. 7). For the 'pass' scan (Fig. 7a) activation is highest just below one of the eyes, but there are also high levels of activation above the forehead. With a complete defacing failure, where behind the face has been defaced and the defacing has gone quite deep into the sides of the head (Fig. 7b), there is very strong activation around a hole in the forehead. Other areas of the head where the defacing has not been adequately applied are also activated. In a scan where the defacing has gone very deep and intersected the brain (Fig. 7c), there is strong activation at the front of the brain and around the strong angular lines where the mask has cut through the sides of the front of the head. Interestingly, only one render shows any activation. The 'shallow' scan, in which the eyes are still visible (Fig. 7d), has very strong activation around the orbits, and high levels around the whole face. In particular the 'L'-shaped cuts where the defacing mask has started to cut into the area of the brows, and below the eyes show strong activation.

Table 1
Confusion matrix showing the CNN's test set performance.

		Actual			
		Deep	Failure	Pass	Shallow
Predicted	Deep	639	99	161	5
	Failure	26	527	5	39
	Pass	126	57	1175	238
	Shallow	9	117	259	1318

Table 2
Confusion matrix showing the CNN's test set performance at binary classification.

		Actual	
		Non-pass	Pass
Predicted	Non-pass	2779	425
	Pass	421	1175

Table 3
CNN performance on the test set, shown for all scans in the set and when subset by modality. Acc = accuracy, Sens = sensitivity, Spec = specificity.

Modality	Images	Acc	Deep		Failure		Pass		Shallow	
			Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
All	4800	0.762	0.799	0.934	0.659	0.983	0.734	0.868	0.824	0.980
T1w	1013	0.762	0.857	0.913	0.494	0.985	0.762	0.844	0.751	0.906
T2w	1256	0.713	0.643	0.921	0.589	0.975	0.535	0.884	0.880	0.808
T1w (gadolinium)	683	0.804	0.861	0.919	0.810	0.977	0.789	0.888	0.717	0.946
PD	981	0.790	0.714	0.988	0.609	0.988	0.817	0.859	0.842	0.832
MT	373	0.753	0.656	0.947	0.824	1.000	0.680	0.809	0.807	0.848
FLAIR	494	0.781	0.854	0.895	0.756	0.973	0.774	0.870	0.759	0.945

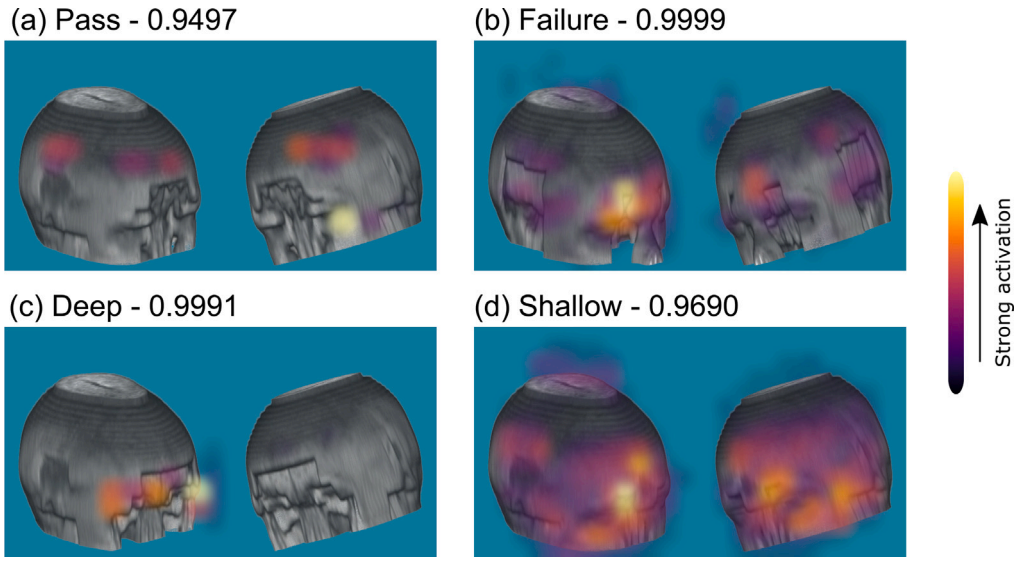

Fig. 7. Grad-CAM heatmaps for representative scans of each of the four classifications, (a) pass, (b) registration failure, (c) deep, and (d) shallow. The CNN probability that each image belongs to the class is shown above each image. Each image was produced from a publicly-available sample scan from [31] with the defacing parameters modified to produce different outcomes.

The CNN is able to classify 1000 images in 51 s (0.051 s per image), including the production of CSV files recording the classifications, and moving images to sub-directories based on assigned classes. When the number of images is increased to 10,000 the processing time decreases to 0.048 s per image (475 s in total).

3.2. Probability threshold selection

As the probability thresholds become less strict, the sensitivity and specificity decline in a similar pattern for the 'pass', 'deep' and 'shallow' classes, although the specificity remains more consistent for the 'deep' class (Fig. 8). While there is a much greater decline in the sensitivity for the 'failure' class compared to the other three classes, the specificity remains very high (>0.97) regardless of the threshold. When strict thresholds (0.95) are applied, the proportion of data which surpasses the selected thresholds is low for all classes (<13%) except the 'failure' class where it is 46%. After examining the overall performance of the network, a global probability threshold of 0.8 was selected for all classes, with the exception of the 'pass' class. As scans assigned the 'pass' classification would undergo a visual verification check a slightly lower threshold of 0.75 was selected. The overall performance of the CNN is greatly improved after applying the thresholds (Table 4). The accuracy has increased to 0.92 (from 0.76 before applying the thresholds), with the sensitivity between 0.91–0.93, and the specificity ranging between 0.94–0.99 across the four classes. Using multi-class ROC curves (Fig. 6b), the AUCs from all comparisons are ≥0.91, with most comparisons having AUCs ≥ 0.93. With the selected thresholds applied, 45% of data in the test set would be classified by the CNN.

Table 4
Confusion matrix showing the CNN's test set performance with the thresholds applied.

		Actual			
		Deep	Failure	Pass	Shallow
Predicted	Deep	302	12	21	0
	Failure	2	388	0	9
	Pass	20	9	682	51
	Shallow	1	18	35	606

Table 5
Confusion matrix showing the CNN's performance at binary classification once the thresholds have been applied.

		Actual	
		Non-pass	Pass
Predicted	Non-pass	1338	56
	Pass	80	682

The CNN with the thresholds applied performs well at identifying 'pass' versus 'non-pass' classifications (Table 5). The accuracy has improved to 0.94, and the sensitivity and specificity have increased to 0.94 and 0.92 respectively, when only the classifications reaching the thresholds are used.

If the thresholded data is broken down by MRI modality (Table 6), the performance is generally consistent between modalities with a small number of exceptions. Accuracy ranges from 0.91 in the T2w and T1w gadolinium scans to 0.96 in the MT scans. The sensitivity is >0.90 for most classes across the modalities, but it still remains low in some

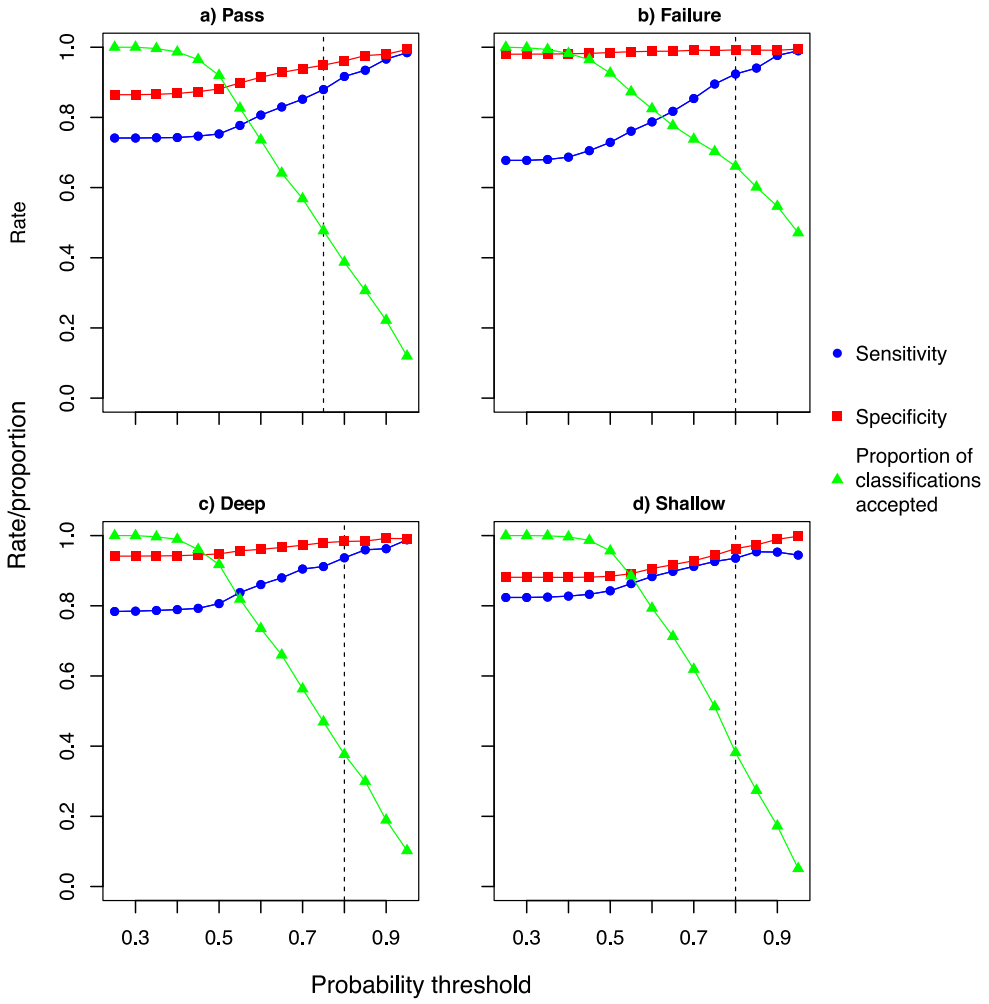

Fig. 8. Plots showing how model performance on the test set varies when the probability threshold of accepted classifications is modified. Each class is shown separately: (a) pass, (b) registration failure, (c) deep, and (d) shallow. Sensitivity, specificity and the proportion of classifications that would be accepted are shown for each class (i.e. scans whose classifications would be discarded at each threshold are not included in the metrics). Vertical dashed lines show the thresholds that were used in the final model.

Table 6
CNN performance on the test set with the thresholds applied. Data is shown for all modalities combined and when subset by modality. Acc = accuracy, Sens = sensitivity, Spec = specificity.

Modality	Images	Acc	Deep		Failure		Pass		Shallow	
			Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
All	2156	0.917	0.929	0.982	0.909	0.994	0.924	0.944	0.910	0.964
T1w	505	0.905	0.944	0.950	0.763	0.994	0.941	0.907	0.809	0.978
T2w	455	0.914	0.789	0.991	0.923	0.985	0.754	0.972	0.963	0.911
T1w (gadolinium)	399	0.915	0.950	0.954	0.924	0.989	0.916	0.955	0.719	0.984
PD	448	0.922	0.864	0.995	0.880	1.000	0.945	0.926	0.911	0.945
MT	155	0.955	0.600	1.000	1.000	1.000	0.973	0.949	0.960	0.982
FLAIR	194	0.923	1.000	0.983	0.936	1.000	0.961	0.937	0.830	0.980

cases despite the application of the thresholds. In some cases these low values are partially the result of the small number of scans in those classes (e.g. only 5 MT scans are classified as 'deep' after applying the thresholds). Specificity is very high in some cases — reaching 1 in the 'failure' class for the three of the modalities. Specificity is lowest (0.91) in the 'pass' class for the T1w modality.

3.3. Model evaluation on a dissimilar dataset

After defacing the IXI MRI scans with the default *fsl_deface* parameters, the manual QC checks showed that the majority of scans had been defaced too 'shallow' (61%), with 36% 'pass' and <2% for each of the 'deep' and 'failure' classes. Of the scans that were defaced using the refined parameter set, 44% of scans were manually classified as 'pass', 34% as 'shallow', 19% as 'deep' and 3% as 'failure'. The CNN achieved 79% accuracy (474 correct) and 77% accuracy (459 correct) for the datasets with the default and refined defacing parameters respectively (Table 7), which is slightly better than the performance of the CNN on the test dataset.

When the selected thresholds were applied, only 212 of the 600 classifications (35%) surpassed the thresholds, of which 191 (90%) of classifications were correct for the scans with the default defacing. Fewer classifications met the thresholds (154; 26%) in the dataset with the refined defacing, and 141/154 (92%) of the classifications matched the manually assigned ones. Whilst the accuracy was similar to that of the test set, the proportion of data that was classified was

Table 7
CNN performance on the IXI dataset for scans defaced using the default and refined parameters. N = number of manually classified scans in the class, Acc = accuracy, Sens = sensitivity, Spec = specificity.

Thresholds applied	Defacing	Acc	Deep			Failure			Pass			Shallow		
			N	Sens	Spec	N	Sens	Spec	N	Sens	Spec	N	Sens	Spec
No	Default	0.790	9	0.444	0.966	10	0.500	0.966	213	0.671	0.915	368	0.875	0.772
No	Refined	0.765	114	0.658	0.957	19	0.579	0.972	265	0.774	0.827	202	0.832	0.884
Yes	Default	0.901	3	0.333	0.995	4	1.000	0.976	68	0.897	0.938	137	0.912	0.920
Yes	Refined	0.916	23	0.826	0.977	8	1.000	0.966	79	0.962	0.933	44	0.864	1.000

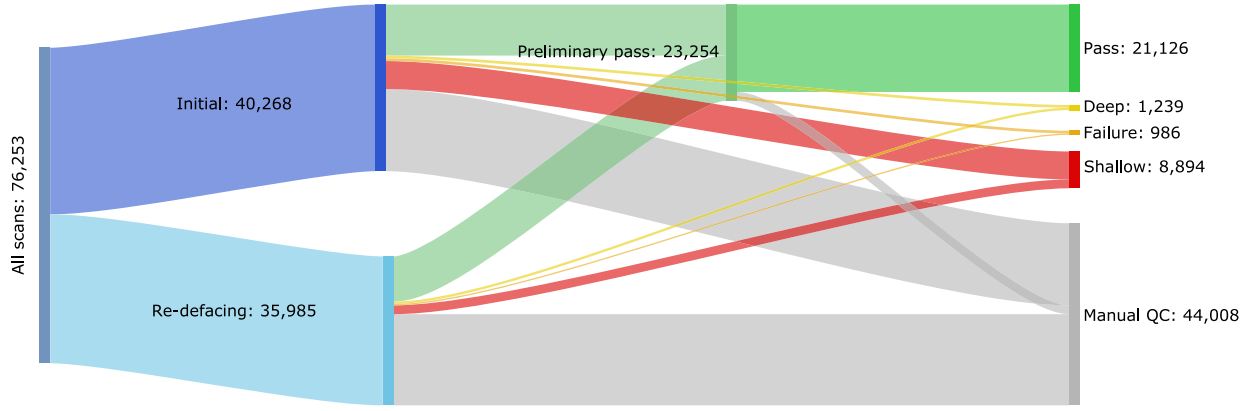

Fig. 9. Sankey diagram showing scans that have been QC checked with the CNN incorporated into the anonymisation pipeline. Scans are split into those from initial defacing and re-defacing QC.

considerably lower. For all classes, across both versions of the IXI data, the specificity was high (≥0.92). In both of the defaced versions of the IXI dataset, some classes only had small numbers of scans (<10) so the specificity and sensitivity are not reliable performance metrics in these cases. Excluding these cases, the sensitivity was typically >0.86 which is noticeably lower than in the study test dataset.

3.4. Integrating the CNN into the defacing pipeline

Since integrating the CNN into the defacing pipeline, over 76,000 scans have been processed (Fig. 9), of which nearly 53% were from scans that had been through initial defacing, and the remaining 47% had been through re-defacing. The CNN classified 30% of scans as passed, 12% as shallow, 2% as deep, and 1% as registration failures; the remaining scans were flagged for manual QC checks as they did not surpass the applied probability thresholds. Of the scans assigned 'pass', 91% and 90% of these classifications were accepted following the visual verification checks in the initial and re-defacing stages respectively. Overall, 45% of the classifications generated by the CNN were accepted, however a larger proportion of the classifications (51%) were accepted for scans that were QC checked following initial defacing, and a much smaller proportion of classifications (39%) were accepted for scans that had been through re-defacing. Of the scans classified by the CNN (i.e. those that did not get flagged for manual QC checks), 60% of these scans were classified as 'pass' during initial defacing, but this was much greater during re-defacing, where 78% of scans were classified as 'pass'. This seems to be the result of a considerably smaller proportion of scans being assigned 'shallow' during re-defacing, due to the success of the revised defacing parameters that are applied during this stage.

Incorporation of the CNN into the pipeline has allowed for a reduction in the amount of time spent on manual QC checks. Using summary data on scans that have been processed so far, it is possible to compute the time that is saved on an average study containing 10,000 scans (Fig. 10). Typically, 31% of scans in a study will need re-defacing, and 12% of the original scans will need a second round of re-defacing. On average, it takes a human scorer 3.8 s to visually QC check a scan, which includes time spent waiting for the image to load on the HTML page, looking at the image, and marking the classification. The visual verification checks are much quicker, taking 0.8 s per scan on average. Therefore, an average study with 10,000 scans would take 15.1 h to manually QC check. With the addition of the CNN to the pipeline the time needed for manual and verification checks during initial defacing would be nearly halved to 6.1 h (from 10.6 h). During re-defacing, the time savings are less due to the CNN not performing as well on re-defaced scans (i.e. these scans are more commonly borderline between classes, or the subject's anatomy or the scan quality makes defacing difficult). Verification and manual QC checks combined would take 2.5 h during the first round and 1.2 h during the second round of re-defacing when using the CNN. Prior to incorporation of the CNN, the manual checks would have taken 3.3 and 1.3 h during the first and second rounds respectively.

In total, QC checks with the CNN incorporated into the pipeline would have taken 9.8 h, instead of 15.1 h prior to incorporation of the CNN into the pipeline, which is a 35% reduction in the time spent performing the checks. Additionally, with the CNN pipeline the entire QC process can be completed in less time than it takes to complete just the initial defacing QC using the original manual pipeline. During just the initial defacing – the most time-consuming part of QC checks – the time spent undertaking QC checks is reduced by 42% using the CNN pipeline. However, the benefits of the CNN are not as substantial during the re-defacing rounds where there is a 23% reduction during the first round, and only a 7% reduction during the second round. If the CNN pipeline were applied to a brand new project of the same size as this one (235,000 scans), the time spent undertaking QC checks could be reduced by approximately 125 h.

4. Discussion

4.1. CNN performance

Prior to applying probability thresholds to the final CNN, the network delivers test accuracy (0.76) which is similar to the performance of models applied to perform other forms of QC on large, multi-site MRI datasets [e.g. 24,27]. However, as the CNN is used to QC check de-identification, higher levels of accuracy were required. The application of the strict probability thresholds conferred much higher test accuracy

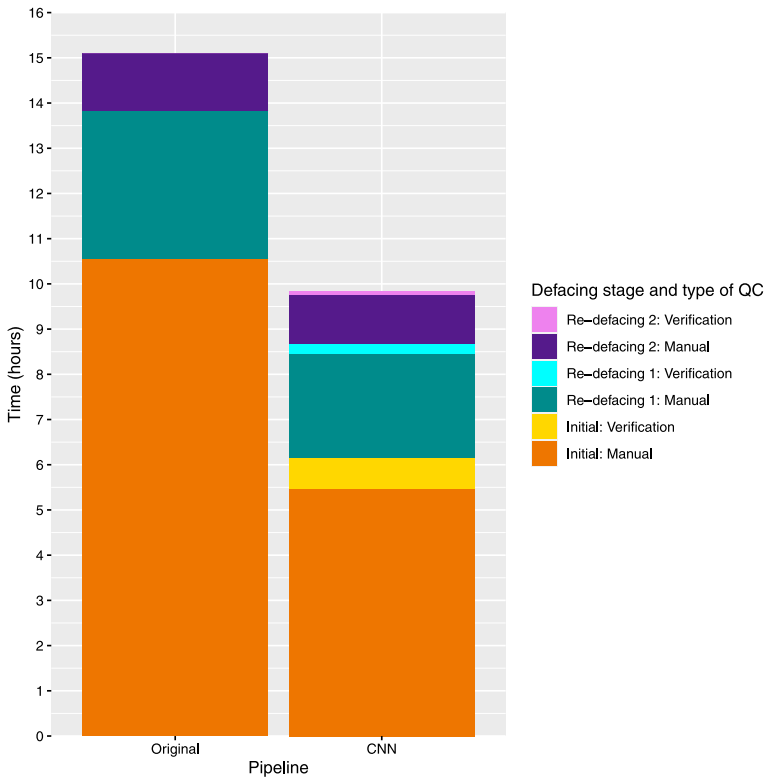

Fig. 10. Stacked bar plot comparing the amount of time spent performing QC checks for an example study containing 10,000 scans, using the original fully manual pipeline and the pipeline with the CNN included.

(0.92) but still allowed for nearly half of the MRI scans to be QC checked without the need for time-consuming manual checks. Our CNN is also able to generalise well across MRI modalities and when applied to externally sourced MRI datasets. Although there was some minor variation, the CNN's performance was relatively consistent across all of the modalities we evaluated it on. This pattern was found both before and after applying the thresholds. Additionally, the CNN performed well on a publicly available dataset, indicating that our model could be applied more widely with comparable results to those demonstrated here.

Choosing appropriate thresholds was key to the successful integration of the CNN into the defacing pipeline. The trade-off between the volume of data processed and classification accuracy had to be considered with relation to the impact of the integration of the CNN into the pipeline. If less strict thresholds were applied, a greater volume of scans could be classified. However, even with the strict thresholds we applied here, up to 10% of assigned 'pass' classifications were not agreed with during verification checks. Further relaxing of the thresholds would most likely lead to a greater proportion of 'pass' assigned scans needing to be flagged for manual checks, increasing the average time to verify each scan, while making the verification checks less efficient and potentially eroding the time savings that the CNN provides. Furthermore, if the probability thresholds for the 'deep', 'failure', and 'shallow' categories were lowered then the quality of re-defacing attempts could be reduced. During the re-defacing, the custom parameters that are applied are highly dependent on the previous QC classifications. Inaccurate 'failure', 'deep' and 'shallow' classifications would lead to incorrectly applied re-defacing parameters, and a greater proportion of scans not passing the re-defacing, and consequently less data would be available for downstream analysis.

To protect patient identities, image anonymisation must remove all identifiable features as the presence of isolated facial features can compromise the anonymisation of image data. For example, a variety of machine learning approaches have been used to develop ear recognition tools, allowing ears to be used for biometric identification [43]. With the application of visualisation methods like Grad-CAM we are able to show that our CNN has been trained to rely on the same anatomical features that human scorers use when deciding whether a scan has been defaced correctly. More precisely, we can verify that the network focuses on the eyes, ears, nose and to a lesser extent the mouth regions (defacing issues around the mouth region are less common). In particular, there was strong activation around the eyes for all classes, which is consistent with observations during the manual QC checks — that the eyes are one of the most common areas to exhibit problems. Also, there are rarely issues with the removal of the ears during the defacing, but it is reassuring that the CNN utilises features around this area of the head. Additionally, the CNN is able to correctly classify scans in the registration failure category despite the appearance of scans in this category being quite variable.

4.2. Comparisons to existing methods

There is no existing comparable baseline automated or machine learning based method for conducting QC checks on the application of defacing, which means that the current gold standard is manual QC checks. In this paper we have extensively evaluated our model against manual checks to validate its performance. Recently, and since developing our own CNN, Bansal et al. [28] have released their *nondefaced detector*, a binary classifier that they developed to identify when T1w scans have (or have not) been defaced with *pydeface* [12]. Bansal et al. [28] report impressive accuracy and sensitivity for their model. In our own supplementary analysis (Supplementary material) using the *nondefaced detector* in similar circumstances to our own model, we found that their model classified 86% of scans that had defacing errors (where the eyes and/or nose had not been successfully removed) as 'defaced'. For comparison, our CNN classified 13% of scans with defacing errors as 'pass' prior to applying the thresholds, and 6% after applying them. If we had included the *nondefaced detector* in our pipeline, a large number of scans would have been marked as correctly

defaced when they actually retained facial features. It should be noted that the model of Bansal et al. [28] is currently published as a preprint and the authors indicate they are continuing to work on developing their model, so it is possible that the future performance of their model will improve in this regard.

It seems probable that the *nondefaced detector* is trained to recognise whether the defacing mask has been applied, rather than the features of the face. For the QC of defacing, being able to specify whether a defacing mask has been applied is not sufficient for guaranteeing de-identification of imaging data. It is important that automated defacing QC models utilise facial features, or the interaction between the defacing mask and those features, as we have been able to show our model is capable of by using Grad-CAM and by validating our model's performance against gold standard manual QC checks. Our choice to use the 3D renders rather than the MRI scans as the input data is a possible reason for the strong performance of our model in this regard. This is because the renders only show the interaction between the defacing mask and the face, and the mask itself is not captured in the render. An additional feature of our model is that it is able to detect when defacing has encroached on the brain. If defacing inadvertently erases parts of the anatomical area of interest, then the data will not be useable by researchers. The major disadvantage of our model compared to the *nondefaced detector*, is that we are not able to publicly share our model as it has been developed using patient data. However, to our knowledge, our CNN is the only model that can detect not just when a defacing mask has been applied to a scan, but it can also identify when its application is incorrect and compromises the de-identification of the scan or impinges on the brain. Furthermore, our model can generalise well across datasets, and is able to handle a number of MRI modalities, not just T1w scans.

4.3. Applicability to other projects

For projects like the Novartis-Oxford collaboration, that utilise hundreds of thousands of MRI scans acquired from sites around the world, one of the main challenges is being able to process the large volumes of data in a timely manner without compromising on the quality of the processing. While machine learning approaches, such as SVMs and CNNs, are regularly used when performing QC checks on MRI data, they are typically used to identify scans that are likely to be problematic when analysed (e.g. poor image quality, identification of artefacts). Our innovative approach of applying a CNN to assist with QC checking the anonymisation of image data highlights that there is potential for applying machine learning approaches to other stages of MRI processing pipelines. The aim of this paper was to develop a model that would assist with the QC of the defacing, and complement, but not replace visual QC checks, while maintaining high QC standards. This goal has been achieved with a reduction in the time spent undertaking manual checks by approximately 35% while maintaining the quality of the images that are passed through to the analytical pipelines. On a project of this scale the time savings are considerable and greatly improve the efficiency of the MRI anonymisation process. Therefore, other projects or platforms dealing with the challenges of anonymising large quantities of MRI data could also find that applying similar approaches as the one detailed in this paper can lead to substantial reductions in the time spent on manual QC processes. Additionally, different defacing methods will likely require bespoke models to capture the relevant features. However, if projects and platforms have existing QC classification data available, then impactful QC models – that are applicable to existing processing pipelines – can be developed even when using relatively simple deep learning architectures. Manual QC checks can be time-consuming bottlenecks, but with the application of approaches such as CNNs, this can be alleviated, expediting the availability of anonymised MRI data to researchers.

4.4. Limitations and future work

Whilst the inclusion of the CNN has made our defacing pipeline more efficient, there is still a continued requirement for manual QC checks. With continued development it may be possible to have a pipeline with minimal or no QC checks. This ambitious goal would require a model with extremely high performance, and there are a number of additional steps which could bring the model closer to performing at this level. While training the model we needed to balance the computational needs for developing the CNN against continuing to run our existing defacing pipeline. If the model was re-developed using more computational resources, then more exhaustive parameter tuning and a more complex network architecture could yield performance improvements. Breaking down the model into two or more separate networks could also be beneficial. For example having an initial model focusing on identifying scans that have passed defacing and an additional network to identify the type of defacing error. Alternatively, using the CNN at different stages of the pipeline could improve efficiency. For example, using the CNN prior to the QC stage of the pipeline to perform automated parameter selection instead of using the default defacing parameters. This could increase the proportion of scans that are defaced correctly during initial defacing, and reduce the number of scans that need re-defacing.

Having our CNN undertake QC using the four different classes was crucial for our pipeline, but presented some challenges. Classifying scans into the four discrete categories is not always straightforward, both for humans and the CNN, as scans often exhibit features characteristic of multiple classifications. For example, the front of the brain is often visible through the forehead in correctly defaced scans, or subjects with deeply-set eyes may show features of 'deep' and 'shallow' classifications. Furthermore, scans are often borderline between two categories (typically 'pass'/'deep' or 'pass'/'shallow'). In these cases, the CNN can be ambiguous with regard to assigning two (or more) classes. A more complex model architecture where the model is split into sub-models with one for each facial feature could improve performance, as each scan could then be classified for the presence of each feature independently. This approach would require a more extensive QC process to develop a training dataset where each scan would be classified for each facial feature.

A limitation of our approach is that as our model has been developed using patient data, we are unable to make it publicly available. Unfortunately, sharing models used for QC checking anonymisation can be problematic when they are developed using patient data, due to the potential for un-anonymised data to be retained in these models [e.g. 44]. Therefore, it is necessary to restrict the storage and usage of these models to a secure environment to protect them from model inversion attacks. Future work could involve re-developing our model using publicly available data, and/or using machine learning approaches such as general adversarial networks (GANs) to create additional MRI training data [e.g. 45] as the volume of publicly available MRI data without defacing already applied is understandably limited.

Our CNN was trained exclusively on scans that had been defaced using *fsl_deface*, and because of this the model is trained to recognise the way that this specific defacing mask interacts with the face. For example, our Grad-CAM analysis showed that there is strong activation around the angular cut marks where the face and the defacing mask interact, and these marks would not be present when defacing masks from other defacing software are used. Therefore, the performance of our model would be sub-optimal for other defacing methods, and it is likely that this pattern would hold true for other automated defacing QC methods if they were developed using only one defacing method. Future defacing QC models could generalise across scans processed by a variety of defacing software by using diverse input data that has been defaced using multiple approaches.

5. Conclusions

In this paper we have developed a CNN that can perform QC checks on the application of MRI defacing. Our model performs well at classifying scans into four classes – one for scans that are correctly defaced and three groups of defacing errors – with 76% accuracy. Furthermore, our model is reliable (82% accuracy) at identifying scans that have been defaced correctly compared to all defacing errors combined as one class. Before integrating our CNN into our pre-existing defacing pipeline we added strict thresholds so ensure that only classifications with high confidence would be accepted. With the addition of these thresholds, the model's accuracy increased to 92% and allowed for around half of the scans to be classified without time-consuming manual checks. With these thresholds, our model was 94% accurate at identifying correctly defaced scans versus scans with erroneous defacing. Our CNN is able to generalise well across MRI modalities, both pre and post application of the thresholds. Additionally we have been able to show that our model can perform well on defaced scans from a publicly available dataset. Implementation of the CNN into the pipeline has lead to a considerable reduction (35%) in the amount of time spent performing manual QC checks, and with future development it is likely that this can be improved further. While we are unable to publicly share our model, our approach is applicable to other similar projects, and has the potential to greatly reduce the burden that manual QC checks can have when verifying the correct de-identification of imaging data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Due to privacy requirements, we are unable to make the data available for sharing. The underlying code used to develop the model in this paper is publicly available (10.5281/zenodo.6638765).

Acknowledgements

The project, and the collaboration, was made possible through access to MRI data from Novartis' MS clinical trials. We wish to thank Piet Arden, Frank Dahlke, and Karine Lheritier from Novartis for their assistance with providing access to data and supporting the collaboration. We wish to acknowledge the help of Mark Jenkinson in providing guidance with establishing the defacing pipeline; Stephen Gardiner, Ewan Straiton, and other members of the data wrangling team for their assistance in working with the MS MRI data; Anna Zalevski and Joanna Stoneham for project management; Adam Huffman, Geoffrey Ferrari, and Robert Esnouf for their work on the IT infrastructure and data transfers for the project.

Funding sources

This paper is the output from the research collaboration between Novartis and the University of Oxford's Big Data Institute. This study was funded by Novartis, Switzerland, and it uses data collected from Novartis funded clinical trials.

Ethics statement

Data used by the collaboration was sourced from Novartis clinical trials, and was approved by institutional review boards or ethics committees. All trials were conducted in accordance with the principles of the Declaration of Helsinki and the International Conference on Harmonisation guidelines for Good Clinical Practice.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106211.

References

[1] A.-M. Mallon, D.A. Häring, F. Dahlke, P. Aarden, S. Afyouni, D. Delbarre, K. El Emam, H. Ganjgahi, S. Gardiner, C.H. Kwok, D.M. West, E. Straiton, S. Haemmerle, A. Huffman, T. Hofmann, L.J. Kelly, P. Krusche, M.-C. Laramee, K. Lheritier, G. Ligozio, A. Readie, L. Santos, T.E. Nichols, J. Branson, C. Holmes, Advancing data science in drug development through an innovative computational framework for data sharing and statistical analysis, BMC Med. Res. Methodol. 21 (1) (2021) 250, http://dx.doi.org/10.1186/s12874-021-01409-4.

[2] F. Dahlke, D.L. Arnold, P. Aarden, H. Ganjgahi, D.A. Häring, J. Čuklina, T.E. Nichols, S. Gardiner, R. Bermel, H. Wiendl, Characterisation of MS phenotypes across the age span using a novel data set integrating 34 clinical trials (NO.MS cohort): Age is a key contributor to presentation, Mult. Sclr. J. 27 (2021) 2062–2076, http://dx.doi.org/10.1177/1352458520988637.

[3] F. Ségonne, A. Dale, E. Busa, M. Glessner, D. Salat, H. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in MRI, NeuroImage 22 (3) (2004) 1060–1075, http://dx.doi.org/10.1016/j.neuroimage.2004.03.032.

[4] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, R.M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (5) (2001) 856–876, http://dx.doi.org/10.1006/nimg.2000.0730.

[5] S.M. Smith, Fast robust automated brain extraction, Hum. Brain Mapp. 17 (3) (2002) 143–155, http://dx.doi.org/10.1002/hbm.10062.

[6] C. Fennema-Notestine, I.B. Ozyurt, C.P. Clark, S. Morris, A. Bischoff-Grethe, M.W. Bondi, T.L. Jernigan, B. Fischl, F. Segonne, D.W. Shattuck, R.M. Leahy, D.E. Rex, A.W. Toga, K.H. Zou, Morphometry BIRN, G.G. Brown, Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location, Hum. Brain Mapp. 27 (2) (2006) 99–113, http://dx.doi.org/10.1002/hbm.20161.

[7] N. Schimke, M. Kuehler, J. Hale, Preserving privacy in structural neuroimages, in: D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, Y. Li (Eds.), Data and Applications Security and Privacy XXV, Vol. 6818, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 301–308, http://dx.doi.org/10.1007/978-3-642-22348-8_26.

[8] N. Schimke, J. Hale, Quickshear defacing for neuroimages, in: HealthSec'11: Proc. 2nd USENIX Conf. Health Secur. Priv., 2011, p. 11.

[9] M. Milchenko, D. Marcus, Obscuring surface anatomy in volumetric imaging data, Neuroinformatics 11 (1) (2013) 65–75, http://dx.doi.org/10.1007/s12021-012-9160-3.

[10] A. Bischoff-Grethe, I.B. Ozyurt, E. Busa, B.T. Quinn, C. Fennema-Notestine, C.P. Clark, S. Morris, M.W. Bondi, T.L. Jernigan, A.M. Dale, G.G. Brown, B. Fischl, A technique for the deidentification of structural brain MR images, Hum. Brain Mapp. 28 (9) (2007) 892–903, http://dx.doi.org/10.1002/hbm.20312.

[11] F. Alfaro-Almagro, M. Jenkinson, N.K. Bangerter, J.L. Andersson, L. Griffanti, G. Douaud, S.N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, D. Vidaurre, M. Webster, P. McCarthy, C. Rorden, A. Daducci, D.C. Alexander, H. Zhang, I. Dragonu, P.M. Matthews, K.L. Miller, S.M. Smith, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank, NeuroImage 166 (2018) 400–424, http://dx.doi.org/10.1016/j.neuroimage.2017.10.034.

[12] O.F. Gulban, D. Nielson, R. Poldrack, J. Lee, C. Gorgolewski, Vanessasaurus, S. Ghosh, Poldracklab/pydeface, 2019, (accessed 17th December 2021). URL https://github.com/poldracklab/pydeface.

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

[14] D.S. Marcus, M.P. Harms, A.Z. Snyder, M. Jenkinson, J.A. Wilson, M.F. Glasser, D.M. Barch, K.A. Archie, G.C. Burgess, M. Ramaratnam, M. Hodge, W. Horton, R. Herrick, T. Olsen, M. McKay, M. House, M. Hileman, E. Reid, J. Harwell, J. Schindler, J.S. Elam, S.W. Curtiss, D.C. Van Essen, W.-M.H. Consortium, Human Connectome Project Informatics: quality control, database services, and data visualization, NeuroImage 80 (2013) 202–219, http://dx.doi.org/10.1016/j.neuroimage.2013.05.077.

[15] A.E. Theyers, M. Zamyadi, M. O'Reilly, R. Bartha, S. Symons, G.M. MacQueen, S. Hassel, J.P. Lerch, E. Anagnostou, R.W. Lam, B.N. Frey, R. Milev, D.J. Müller, S.H. Kennedy, C.J.M. Scott, S.C. Strother, Multisite comparison of MRI defacing software across multiple cohorts, Front. Psychiatry 12 (2021) 617997, http://dx.doi.org/10.3389/fpsyt.2021.617997.

[16] G.V. Bhalerao, P. Parekh, J. Saini, G. Venkatasubramanian, J.P. John, B. Viswanath, N.P. Rao, J.C. Narayanaswamy, P.T. Sivakumar, A. Kandasamy, M. Kesavan, U.M. Mehta, O. Mukherjee, M. Purushottam, R. Kannan, B. Mehta, T. Kandavel, B. Binukumar, D. Jayarajan, A. Shyamsundar, S. Moirangthem, K. Vijay Kumar, J. Mahadevan, B. Holla, J. Thirthalli, P.S. Chandra, B.N. Gangadhar,

P. Murthy, M.M. Panicker, U.S. Bhalla, S. Chattarji, V. Benegal, M. Varghese, J.Y. Reddy, P. Raghu, M. Rao, S. Jain, Systematic evaluation of the impact of defacing on quality and volumetric assessments on T1-weighted MR-images, J. Neuroradiol. 49 (2022) 250–257, http://dx.doi.org/10.1016/j.neurad.2021.03.001.

[17] C.G. Schwarz, W.K. Kremers, H.J. Wiste, J.L. Gunter, P. Vemuri, A.J. Spychalla, K. Kantarci, A.P. Schultz, R.A. Sperling, D.S. Knopman, R.C. Petersen, C.R. Jack, Changing the face of neuroimaging research: Comparing a new MRI defacing technique with popular alternatives, NeuroImage 231 (2021) 117845, http://dx.doi.org/10.1016/j.neuroimage.2021.117845.

[18] F.W. Prior, B. Brunsden, C. Hildebolt, T.S. Nolan, M. Pringle, S.N. Vaishnavi, L.J. Larson-Prior, Facial recognition from volume-rendered magnetic resonance imaging data, IEEE Trans. Inf. Technol. Biomed. 13 (1) (2009) 5–9, http://dx.doi.org/10.1109/TITB.2008.2003335.

[19] J.C. Mazura, K. Juluru, J.J. Chen, T.A. Morgan, M. John, E.L. Siegel, Facial recognition software success rates for the identification of 3D surface reconstructed facial images: Implications for patient privacy and security, J. Digit. Imaging 25 (3) (2012) 347–351, http://dx.doi.org/10.1007/s10278-011-9429-3.

[20] C.G. Schwarz, W.K. Kremers, T.M. Therneau, R.R. Sharp, J.L. Gunter, P. Vemuri, A. Arani, A.J. Spychalla, K. Kantarci, D.S. Knopman, R.C. Petersen, C.R. Jack, Identification of anonymous MRI research participants with face-recognition software, New Engl. J. Med. 381 (2019) 1684–1686, http://dx.doi.org/10.1056/NEJMc1908881.

[21] D. Abramian, A. Eklund, Refacing: reconstructing anonymized facial features using GANs, in: IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 1104–1108, http://dx.doi.org/10.1109/ISBI.2019.8759515.

[22] A. Keshavan, J.D. Yeatman, A. Rokem, Combining citizen science and deep learning to amplify expertise in neuroimaging, Front. Neuroinform. 13 (2019) http://dx.doi.org/10.3389/fninf.2019.00029.

[23] B. Mortamet, M.A. Bernstein, C.R. Jack, J.L. Gunter, C. Ward, P.J. Britson, R. Meuli, J.-P. Thiran, G. Krueger, Automatic quality assessment in structural brain magnetic resonance imaging: Automatic QA in Structural Brain MRI, Magn. Reson. Med. 62 (2) (2009) 365–372, http://dx.doi.org/10.1002/mrm.21992.

[24] R.A. Pizarro, X. Cheng, A. Barnett, H. Lemaitre, B.A. Verchinski, A.L. Goldman, E. Xiao, Q. Luo, K.F. Berman, J.H. Callicott, D.R. Weinberger, V.S. Mattay, Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm, Front. Neuroinform. 10 (2016) http://dx.doi.org/10.3389/fninf.2016.00052.

[25] E.T. Klapwijk, F. van de Kamp, M. van der Meulen, S. Peters, L.M. Wierenga, Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data, NeuroImage 189 (2019) 116–129, http://dx.doi.org/10.1016/j.neuroimage.2019.01.014.

[26] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.

[27] O. Esteban, D. Birman, M. Schaer, O.O. Koyejo, R.A. Poldrack, K.J. Gorgolewski, MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites, PLoS ONE 12 (9) (2017) e0184661, http://dx.doi.org/10.1371/journal.pone.0184661.

[28] S. Bansal, A. Kori, W. Zulfikar, J. Wexler, C.J. Markiewicz, F.F. Feingold, R.A. Poldrack, O. Esteban, High-sensitivity detection of facial features on MRI brain scans with a convolutional network, BioRxiv (2021) http://dx.doi.org/10.1101/2021.04.25.441373.

[29] Y. Halchenko, M. Goncalves, V.d.O. Castello, S. Ghosh, M. Hanke, M. Brett, J. Carlin, Nipy/heudiconv: Heudiconv, 2019, (accessed 17th December 2021). URL https://github.com/nipy/heudiconv.

[30] K.J. Gorgolewski, T. Auer, V.D. Calhoun, R.C. Craddock, S. Das, E.P. Duff, G. Flandin, S.S. Ghosh, T. Glatard, Y.O. Halchenko, D.A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B.N. Nichols, T.E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J.A. Turner, G. Varoquaux, R.A. Poldrack, The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments, Sci. Data 3 (1) (2016) 160044, http://dx.doi.org/10.1038/sdata.2016.44.

[31] I Do Imaging, I do imaging, 2017, (accessed 17th December 2021). URL https://idoimaging.com.

[32] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, FSL, NeuroImage 62 (2) (2012) 782–790, http://dx.doi.org/10.1016/j.neuroimage.2011.09.015.

[33] J.J. Allaire, F. Chollet, Keras: R interface to 'Keras', 2018, (accessed 17th December 2021). URL https://CRAN.R-project.org/package=keras.

[34] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2018, (accessed 17th December 2021). URL https://www.R-project.org.

[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, http://dx.doi.org/10.48550/ARXIV.1409.1556, arXiv.

[36] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 25, Curran Associates, Inc., 2012, pp. 1106–1114, http://dx.doi.org/10.1145/3065386.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.

[38] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE Int. Conf. Comput. Vis. (ICCV), IEEE, Venice, 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[39] F. Chollet, J.J. Allaire, Deep Learning with R, Manning, Shelter Island, NY, 2018.

[40] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2001) 171–186, http://dx.doi.org/10.1023/A:1010920819831.

[41] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics 12 (1) (2011) 77, http://dx.doi.org/10.1186/1471-2105-12-77.

[42] IXI, IXI - Information eXtraction from Images (EPSRC GR/S21533/02), 2005, (accessed 12th May 2022). URL https://brain-development.org/ixi-dataset/.

[43] Ž. Emeršič, V. Štruc, P. Peer, Ear recognition: More than a survey, Neurocomputing 255 (2017) 26–39, http://dx.doi.org/10.1016/j.neucom.2016.08.139.

[44] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proc. 22nd ACM SIGSAC Conf. Comput. and Commun. Secur., ACM, Denver Colorado USA, 2015, pp. 1322–1333, http://dx.doi.org/10.1145/2810103.2813677.

[45] K. Kazuhiro, R.A. Werner, F. Toriumi, M.S. Javadi, M.G. Pomper, L.B. Solnes, F. Verde, T. Higuchi, S.P. Rowe, Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images, Tomography 4 (4) (2018) 159–163, http://dx.doi.org/10.18383/j.tom.2018.00042.