University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

01 University of Plymouth Research Outputs University of Plymouth Research Outputs

2024-01-12

Deep Learning Detection of Aneurysm Clips for Magnetic Resonance Imaging Safety

Courtman, M

https://pearl.plymouth.ac.uk/handle/10026.1/21636

10.1007/s10278-023-00932-8
Journal of Digital Imaging
Springer

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Journal of Digital Imaging

Deep learning detection of aneurysm clips for magnetic resonance imaging safety

--Manuscript Draft--

Manuscript Number:	JDIM-D-23-00440R1
Full Title:	Deep learning detection of aneurysm clips for magnetic resonance imaging safety
Article Type:	Original Paper
Section/Category:	Technical Imaging Informatics
Funding Information:	Engineering and Physical Sciences Research Council (EP/T518153/1) — Mrs Megan Courtman National Institute for Health and Care Research (PEN/006/005/A) — Dr Mark Thurston
Abstract:	Flagging the presence of metal devices before a head MRI scan is essential to allow appropriate safety checks. There is an unmet need for an automated system which can flag aneurysm clips prior to MRI appointments. We assess the accuracy with which a machine learning model can classify the presence or absence of an aneurysm clip on CT images. A total of 280 CT head scans were collected, 140 with aneurysm clips visible and 140 without. The data were used to retrain a pre-trained image classification neural network to classify CT localizer images. Models were developed using five-fold cross-validation and then tested on a holdout test set. A mean sensitivity of 100% and a mean accuracy of 82% were achieved. Predictions were explained using SHapley Additive exPlanations (SHAP), which highlighted that appropriate regions of interest were informing the models. Models were also trained from scratch to classify three-dimensional CT head scans. These did not exceed the sensitivity of the localizer models. This work illustrates an application of computer vision image classification to enhance current processes and improve patient safety.
Corresponding Author:	Megan Courtman University of Plymouth UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Plymouth
Corresponding Author's Secondary Institution:	
First Author:	Megan Courtman
First Author Secondary Information:	
Order of Authors:	Megan Courtman
	Daniel Kim
	Huub Wit
	Hongrui Wang
	Lingfen Sun
	Emmanuel Ifeachor
	Stephen Mullin
	Mark Thurston
Order of Authors Secondary Information:	
Author Comments:	

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Click here to access/download;Blinded Manuscript;sn-article.tex

Click here to view linked References

Deep learning detection of aneurysm clips for magnetic resonance imaging safety

Abstract

Flagging the presence of metal devices before a head MRI scan is essential to allow appropriate safety checks. There is an unmet need for an automated system which can flag aneurysm clips prior to MRI appointments. We assess the accuracy with which a machine learning model can classify the presence or absence of an aneurysm clip on CT images. A total of 280 CT head scans were collected, 140 with aneurysm clips visible and 140 without. The data were used to retrain a pre-trained image classification neural network to classify CT localizer images. Models were developed using five-fold cross-validation and then tested on a holdout test set. A mean sensitivity of 100% and a mean accuracy of 82% were achieved. Predictions were explained using SHapley Additive exPlanations (SHAP), which highlighted that appropriate regions of interest were informing the models. Models were also trained from scratch to classify three-dimensional CT head scans. These did not exceed the sensitivity of the localizer models. This work illustrates an application of computer vision image classification to enhance current processes and improve patient safety.

Keywords: Aneurysm clips, artificial intelligence, CT, deep learning, MRI, patient safety

1 Introduction

Screening of patients for aneurysm clips and other metallic devices prior to magnetic resonance imaging (MRI) is vital to ensure that the patient and device can be scanned safely. There have been numerous makes and designs of aneurysm clip over decades [1], many of which have been categorized as MRI safe. For these particular implants, MRI is not absolutely contraindicated, but the devices need careful prior assessment to ensure that the scan takes place under manufacturer-specified conditions. However, not all historic clips are MRI safe, and even those that are safe in some conditions may not be safe in all conditions [2]. At least one fatality has been caused by the displacement of an aneurysm clip [3]. Safe examination requires review of medical records and co-ordination of multiple experts [4]. Late detection has the potential to result in last minute cancellations and wasted scanner time. Failure to perform the required checks can result in device dysfunction with potential harm to the patient.

MRI is the standard imaging modality for many conditions. Appropriate screening policies and procedures are essential before permitting entry to the MRI scanner to prevent injury [5]. Best practice is to use referrer and patient questionnaires to identify patients with devices or other issues that need further investigation. Questionnaires are not fail-safe as referrer responses can be unreliable and patient responses are often not available until the day of the scan.

In the last decade, there have been significant advances in AI-based medical image classification due to increased compute power, the open-sourcing of large labelled datasets, and the development of deep learning [6]. Deep learning describes the subset of machine learning which uses layered neural networks to build representations of complicated concepts out of simpler concepts [7]. This negates the need for feature extraction, as required by other methods, and streamlines the preprocessing pipeline [8]. The success of deep learning methods in image classification tasks is well-documented, and for the last decade they have exceeded the performance of many other state-of-the-art classification algorithms [9]. There are now thousands of publications applying deep learning techniques to medical imaging [10].

We describe the design of a deep learning model for the detection of the presence of aneurysm clips in computerized tomography (CT) head scans. The vast majority of patients with aneurysm clips will have had CT head imaging previously performed as part of their treatment, presenting the potential to screen these previous scans as part of an automatic pre-MRI safety check. This would improve MRI safety, reduce last-minute cancellations, and save time and resources.

2 Materials and Methods

Ethical approval was granted on 15 October 2019 by HRA and Health and Care Research Wales. Data were obtained from Derriford Hospital, a large teaching hospital with a regional neurosurgery centre serving the South West of the United Kingdom. The study design was retrospective and observational using pre-existing medical image data.

2.1 Subject Inclusion

A database of patients with aneurysm clips was used to identify cases for inclusion in the study. A list of all patients undergoing aneurysm clip surgery was identified from surgical records. The radiology information system (RIS) (Cris, Wellbeing Software) was used to identify all post-surgical CT head examinations for these patients. A custom SQL query was then used to search the RIS for matched controls. For each scan with an aneurysm clip present, a scan with no aneurysm clip present was identified. These control scans were matched according to:

- scan type
- age at time of scan, within a window of ± six months
- scan date, within a window of ± 12 months
- gender

2.2 Image Data Acquisition

Images for the investigations identified on the RIS were downloaded from PACS using dcmtk (OFFIS e.V.) [11]. These studies were anonymised using custom anonymisation software based on the Clinical Trials Processor (RSNA MIRC project) [12].

2.3 Ground Truth Confirmation

Manual review of images was performed by two board-certified radiologists to ensure correct labelling. In the event of any disagreement of the correct labels, a third board-certified radiologist reviewed the case to confirm the correct labelling.

2.4 Split

Two sets of images were extracted from the fully curated dataset: a set of localizers and a set of full CT heads. Most CT scan studies begin with one or more localizer scans. These are of poorer quality than full CT scans, but aneurysm clips can often still be clearly seen (Fig. 1). Localizer scans acquired in the same plane were identified automatically using the DICOM tags. From the fully curated dataset, 274 scans were identified which contained saggital localizers: 136 with aneurysm clips and 134 without. These localizers were randomly divided at a scan level: 28 scans (10%) were reserved as a holdout test set (10 with aneurysm clips and 18 without). The remaining 246 (90%) were used for model development (126 with aneurysm clips and 120 without).

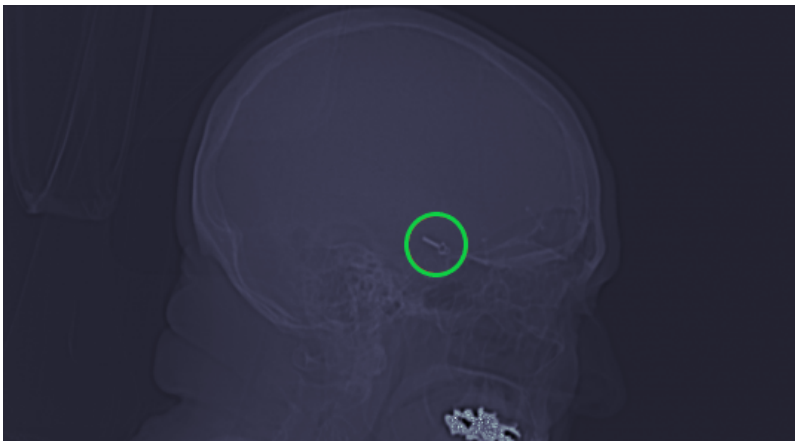


Fig. 1: Sagittal localizer with aneurysm clip present, circled

To standardise the full CT head dataset, scans reconstructed using the same kernel were identified automatically using the DICOM tags. From the fully curated dataset, 214 scans were identified which had been reconstructed using a bone kernel: 104 with aneurysm clips and 110 without. These were randomly divided at a scan level: 22 scans (10%) were reserved as a holdout test set (11 with aneurysm clips and 11 without).

The remaining 192 (90%) were used for model development (93 with aneurysm clips and 99 without).

For both localizers and full CT heads, five-fold cross-validation was used to develop and assess models, with the data divided into 80% training data and 20% validation data in each fold.

For both types of image, the five developed models were then finally tested on the holdout test set.

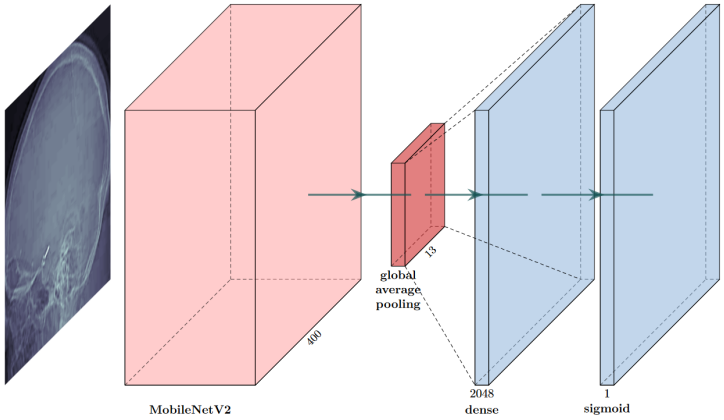
2.5 Image Preprocessing

The images were preprocessed before model input by a deterministic automatic pipeline developed in Python using tools from OpenCV [13], SciPy [14] and scikit-image [15]. For the two-dimensional localizer scans, black borders were removed. Pixel values were rescaled between zero and one. Images were cropped to contain the head only, and the bottom of the images removed to exclude the mandible. This optimisation was included after the explainability technique revealed that models were being confounded by the presence of fillings, resulting in false positive results. Images were resized to 400x400 pixels.

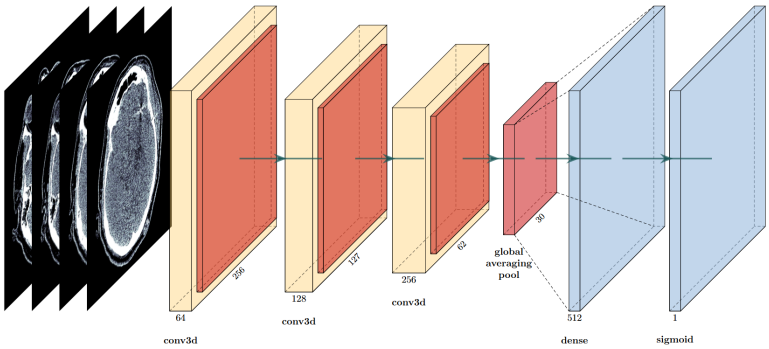
For the three-dimensional scans, the Hounsfield values were clipped with a level of 2000 and a window of 500 to optimize the visibility of metal. Voxel values were scaled between zero and one. Images were cropped to contain the head only and resized to 256x256x40 voxels.

2.6 Neural Network Architecture

Python-based deep neural networks were built with Keras [16] using the TensorFlow backend [17]. Graphics processing unit hardware acceleration on an NVIDIA GeForce RTX 3080 was used for neural network training. Jupyter Lab [18] was used for model development to enable iterative improvements to be made efficiently.



(a) MobileNet V2 for localizers



(b) 3D-CNN for full CT heads

Fig. 2: Network architectures

For the classification of the two-dimensional localizer images, a convolutional neural network based on a pre-trained model was selected as a proven choice for computer vision and image classification tasks using transfer learning [10]. Several well-established pre-trained base networks were trialled, including VGG16 [19], Inception V3 [20], Xception [21], DenseNet [22] and MobileNet V2 [23]. Following analysis for each model, MobileNet V2 achieved the greatest performance and was chosen for the final models (Fig. 2a).

For the classification of the three-dimensional CT images, a three-dimensional convolutional neural network was trained from scratch, due to a lack of available pre-trained three-dimensional classification networks [24]. Several different hyperparameter configurations were trialled. Following curve analysis for each iteration, the one which achieved the smallest loss on the validation data was chosen for the final models (Fig. 2b).

2.7 Model Training

The models were trained for a maximum of 100 epochs using stochastic gradient descent with the Adam optimization algorithm (learning rate 0.001) [25]. The binary cross-entropy loss function was utilized. The batch size was 64. The images were augmented with a 50% probability of horizontal flip. Other augmentation methods were trialled, but did not result in any further increase in performance. The models achieving the lowest loss on the validation sets during training were saved using checkpoints.

A classification threshold was then chosen for the models which maximized sensitivity, and therefore minimized the prevalence of false negatives.

2.8 Explainability

SHapley Additive exPlanations (SHAP) were used to explain the 2D models' predictions. SHAP uses the game theory concept of Shapley values to calculate the contribution of a factor to a machine learning model output [26]. In this case, DeepSHAP was used to calculate and visualize the contribution of individual pixels to the deep learning model's prediction.

3 Results

3.1 Localizer Images

Of the pre-trained base models trialled for the localizer images, MobileNet V2 achieved the greatest mean test Receiver Operating Characteristic (ROC) area under the curve (AUC) and was chosen for the final models. Other base model results are reported in Table 1.

Table 1: Performance of different base models for localizer images

Base model	Mean ROC AUC	Parameters	Inference time (ms)	GFLOPS
VGG16	0.84	15,767,361	24.9	97.9
Inception V3	0.95	26,001,185	27.4	21.0
XCeption	0.98	25,059,881	25.5	29.4
DenseNet	0.98	22,258,241	30.7	27.4
MobileNet V2	0.99	4,883,521	26.2	2.0

A classification threshold of 0.16 was chosen to maximize sensitivity whilst maintaining a high accuracy and specificity (Fig. 3). The final models achieved a mean test sensitivity of 100%. Other performance metrics are reported in Table 2.

When tested on the holdout test set of 28 localizer images, the final models achieved a sensitivity of 100%. Other performance metrics are reported in Table 2.

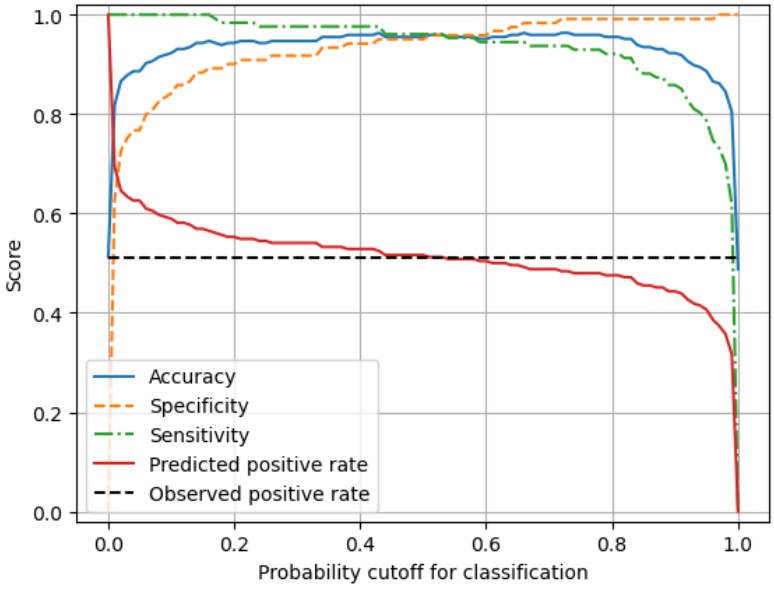


Fig. 3: Mean test performance metrics for MobileNet V2 models in training

Table 2: Performance metrics for MobileNet V2 models with classification threshold of 0.16

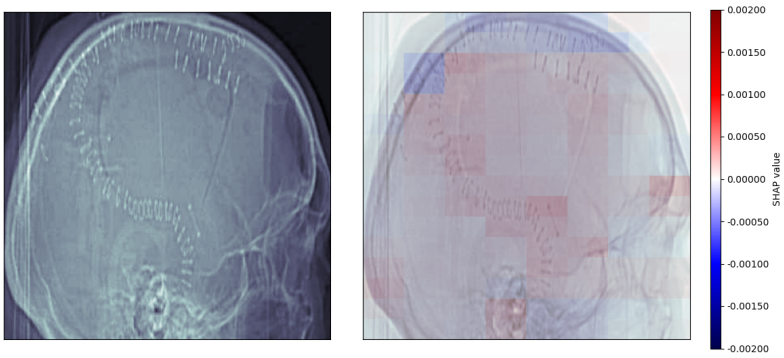
Performance metric	Training mean	Holdout mean
ROC AUC	0.99	1.00
Accuracy	95%	82%
Sensitivity	100%	100%
Specificity	89%	82%

3.1.1 Incorrectly Classified Examples

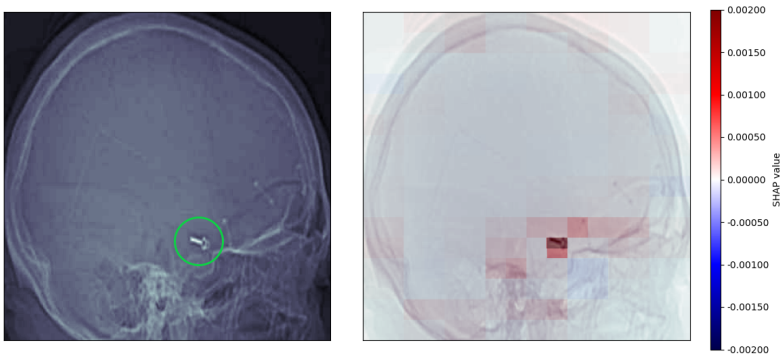
The incorrectly classified 2D localizer images were analysed using the SHAP explainability method. In the early stages of the research, this demonstrated the need to remove the mandible from the images, as prior to this removal the models were confounded by the presence of fillings.

After the images had been cropped and models developed, the SHAP explainability method was used to analyse the incorrectly classified examples in the holdout test set. Three of the 28 images were incorrectly classified by all five models, and five other images were misclassified by at least one of the models. All of these errors were false positives. The average SHAP maps show that bright areas have contributed to the models' incorrect predictions, including other metal devices (Fig. 4a).[1]

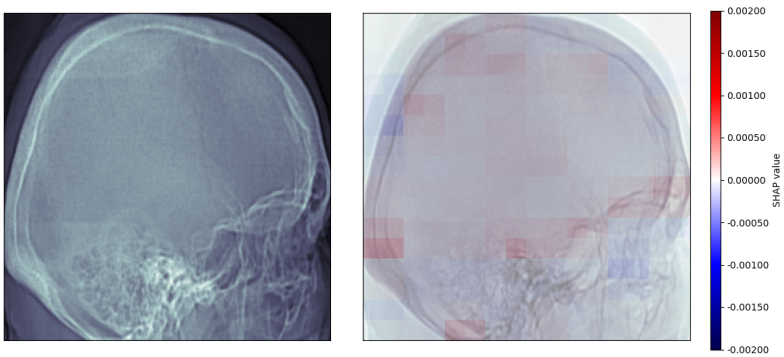
[1]See supplementary figure 1 for all false positive average SHAP maps.



(a) False positive, as predicted by five models. The mean output probability of the image containing a clip is 0.46.



(b) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.99.



(c) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.

Fig. 4: Maps of average SHAP values. Any pixels highlighted in red have contributed to the prediction that an aneurysm clip is present; any pixels highlighted in blue have contributed to the prediction that no aneurysm clip is present. In the case of the true positive, the aneurysm clip has been circled in green for clarity.

3.1.2 Correctly Classified Examples

The SHAP explainability method was also used to analyse the localizer images that the models classified correctly. Of the 28 images in the holdout test set, 20 were classified correctly by all five models. The average SHAP maps for the true positives show that the pixels containing aneurysm clips contributed positively to models' correct predictions that a clip is present (Fig. 4b). [2] The signal is much stronger than the confounding signals in the false positive predictions, and is much stronger than any signal in the true negative predictions where no clip has been detected (Fig. 4c). [3]

3.2 Three-Dimensional CT Images

After models had been trained on three-dimensional CT images, a classification threshold of 0.30 was chosen to maximize sensitivity whilst maintaining a high accuracy and specificity (Fig. 5). The final models achieved a mean test sensitivity of 96%. Other performance metrics are reported in Table 3.

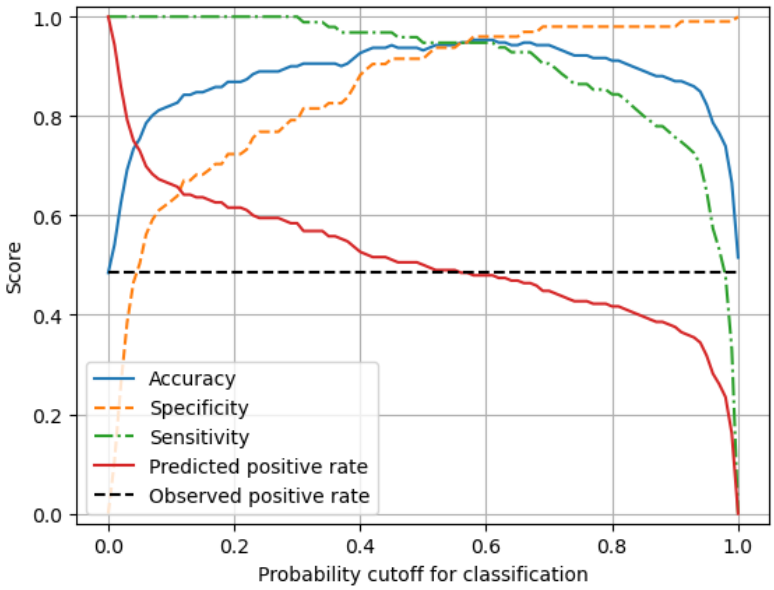


Fig. 5: Mean test performance metrics for 3D models in training

When tested on the holdout test set of 22 three-dimensional CT images, the final models achieved a mean sensitivity of 96%. Other performance metrics are reported in Table 3. Of the 22 images, 19 were correctly classified by all five models. Of the three images that were incorrectly classified by at least one model, two were false positives and one was a false negative.

[2] See supplementary figure 2 for all true positive average SHAP maps.
[3] See supplementary figure 3 for all true negative average SHAP maps.

Table 3: Performance metrics for 3D models with classification threshold of 0.30

Performance metric	Training mean	Holdout mean
ROC AUC	0.99	0.96
Accuracy	90%	95%
Sensitivity	100%	96%
Specificity	79%	95%

4 Discussion

Deep learning has previously been used successfully to detect medical implants. Pre-trained convolutional neural networks have been used to detect pacemakers in chest radiographs with 99.67% accuracy [27] and spinal implants in lumbar spine lateral radiographs with 98.7% precision and 98.2% recall [28]. A convolutional neural network trained from scratch has been used to identify dental implants in X-ray images with 94.0% segmentation accuracy and 71.7% classification accuracy [29]. In another application, a segmentation network has been developed to identify orthopedic implants in hip and knee radiographs with 98.9% accuracy and 100% top-three accuracy, exceeding the performance of five senior orthopedic specialists [30].

The successful implementation of deep learning for implant detection is continued in this application, the first to use deep learning to detect aneurysm clips. The trained models exhibit excellent performance for both localizer images and full CT head scans. Both types of model generalize well to the unseen data in the holdout sets and score particularly highly in terms of sensitivity. The sensitivity for the localizer models is 100% in both the training and the holdout data: there are no dangerous false negatives. The computational resources required to run the models are particularly low in the case of the localizer images.

The use of an explainability method is particularly valuable in this application because it demonstrates that the correct parts of the localizer image are informing the models. In general, the positive (red) signal in the images is strongly localized and more observable than the negative (blue) signal, which is weaker and more distributed. This suggests that the models are being positively informed by the presence of aneurysm clips, and are being informed on a more widespread and low level by the absence of aneurysm clips.

As this application is a potential safety tool, the models have been developed and classification thresholds chosen to maximize sensitivity and minimize false negatives. As a result, they are sometimes confounded by other bright areas in the images, making some false positives likely. This could create additional work for a human operator, but it is a preferable error to dangerous false negatives. The heatmaps also demonstrate that other metal devices such as skull flap fixing plates and skin clips can be responsible for false positives (see supplementary figure 1). These are still valuable to detect for MRI safety. Future work could assess these models on a CT head dataset incorporating a wider range of metallic implants, to analyse whether models trained to detect aneurysm clips specifically generalize to metal implant detection more broadly.

It was anticipated that models developed for full CT heads might perform better than models developed for localizer scans, as the aneurysm clip would be presented in three dimensions and in greater detail. However, the sensitivity of the three-dimensional models was slightly poorer. This may have been due to the presence of too much other confounding detail, or may have been due to the models having been trained from scratch rather than taking advantage of pre-learned patterns. Pre-trained networks were used for the localizer scans due to their ready availability for transfer learning in two-dimensional image data. At this time, there is a notable lack of equivalent pre-trained networks available for transfer learning in three-dimensional image data. If pre-trained three-dimensional networks become available in the future, then they might be successfully leveraged in this application.

Future work could consider using an ensemble model. Ensemble methods are considered the state of the art for many machine learning applications, as they harness the power of weaker learners [31]. An ensemble model for this application could incorporate different learning algorithms, as well as bagging or boosting approaches.

4.1 Limitations

The size of the data is a limitation of this research, caused by the rarity of CT scans depicting aneurysm clips. If it were possible to obtain more data this might enable the development of even more accurate models in training, and enable more representative assessment of models in the holdout set. We have mitigated this limitation to an extent by augmenting the training data with horizontal flip, thus artificially increasing the size of the dataset.

Another limitation of this research is the lack of external validation. External validation sets are difficult to obtain as appropriate publicly available databases do not exist. Our research team is in the process of planning and gaining governance clearance for such accessible studies. We have mitigated this limitation as far as possible in this study by reserving an unseen holdout test set. However, these data originate from the same source as the training data, and the metrics reported may not be representative of the models' performance on data from a different distribution. For example, the balance of the data used in this study is not representative of the typical MRI patient population, in which only a small minority would have aneurysm clips present. An external validation set would allow for more accurate assessment of the models' capability to generalize to other populations.

5 Conclusion

A pre-trained MobileNet V2 neural network achieved high accuracy and 100% sensitivity for the detection of aneurysm clips in CT localizer scans, and the explainability method demonstrated that the network was focusing on appropriate regions of interest in the images. A trained-from-scratch neural network also achieved high accuracy and sensitivity for the detection of aneurysm clips in full CT head scans. This application could be a useful addition to current processes, enabling automatic safety screening for devices in advance of MRI appointments.

References

[1] J. T. McFadden, "Magnetic resonance imaging and aneurysm clips: a review," *Journal of neurosurgery*, vol. 117, no. 1, pp. 1–11, 2012.

[2] M. F. Dempsey, B. Condon, and D. M. Hadley, "MRI safety review," in *Seminars in Ultrasound, CT and MRI*, 2002, vol. 23, no. 5, pp. 392–401.

[3] R. P. Klucznik, D. A. Carrier, R. Pyka, and R. W. Haid, "Placement of a ferromagnetic intracerebral aneurysm clip in a magnetic field with a fatal outcome.," *Radiology*, vol. 187, no. 3, pp. 855–856, 1993.

[4] A. Cunqueiro, M. Lipton, R. Dym, V. Jain, J. Sterman, and M. Scheinfeld, "Performing MRI on patients with MRI-conditional and non-conditional cardiac implantable electronic devices: an update for radiologists," *Clinical Radiology*, vol. 74, no. 12, pp. 912–917, 2019.

[5] F. G. Shellock and A. Spinazzi, "MRI safety update 2008: part 2, screening patients for MRI.," *American Journal of Roentgenology*, vol. 191, no. 4, p. 1140, 2008.

[6] A. Esteva *et al.*, "Deep learning-enabled medical computer vision," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[8] Y. Bengio *et al.*, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[10] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[11] OFFIS, "DCMTK," available via https://dicom.offis.de/dcmtk/. Accessed October 2023.

[12] Radiological Society of North America, Inc., "CTP - The RSNA Clinical Trial Processor," available via https://mircwiki.rsna.org/index.php?title=MIRC_CTP. Accessed October 2023.

[13] G. Bradski, and A. Kaehler, "OpenCV," *Dr. Dobb's journal of software tools*, 3(2), 2000.

[14] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature methods*, 17(3), pp.261-272,2020.

[15] S. Van der Walt *et al.*, "scikit-image: image processing in Python," *PeerJ*, 2, p.e453, 2014.

[16] F. Chollet and others, "Keras." 2015.

[17] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." 2015.

[18] T. Kluyver *et al.*, "Jupyter Notebooks – a publishing format for reproducible computational workflows," in Positioning and Power in Academic Publishing: Players, Agents and Agendas, 2016, pp. 87–90.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[24] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," arXiv preprint arXiv:1904.00625, 2019.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[26] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems* 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp.

[27] M. D. V. Thurston, D. H. Kim, and H. K. Wit, "Neural network detection of pacemakers for MRI safety," *Journal of Digital Imaging*, vol. 35, no. 6, pp. 1673–1680, 2022.

[28] H.-S. Yang, K.-R. Kim, S. Kim, and J.-Y. Park, "Deep learning application in spinal implant identification," *Spine*, vol. 46, no. 5, pp. E318–E324, 2021.

[29] A. Kohlakala, J. Coetzer, J. Bertels, and D. Vandermeulen, "Deep learning-based dental implant recognition using synthetic X-ray images," *Medical & Biological Engineering & Computing*, vol. 60, no. 10, pp. 2951–2968, 2022.

[30] R. Patel, E. H. Thong, V. Batta, A. A. Bharath, D. Francis, and J. Howard, "Automated identification of orthopedic implants on radiographs using deep learning," *Radiology: Artificial Intelligence*, vol. 3, no. 4, 2021.4765–4774.

[31] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

Click here to access/download
Supplementary Material
Aneurysm_clips_supplementary_material-3.pdf

"Deep learning detection of aneurysm clips for magnetic resonance imaging safety" – response to editor

We would like to thank the editor and reviewers for their thorough and helpful comments on the paper. We have subsequently made major revisions. Please find responses to suggestions highlighted below.

Comments to the author:

1) page 2 line 28: Most = what %?

We agree that this would be a useful inclusion – unfortunately precise numbers are difficult to acquire. For clarity, we have updated the wording to "the vast majority of patients with aneurysm clips will have had CT head imaging previously performed as part of their treatment".

2) page 3 line 26: Of the 229 how many had clips vs not? Then for thew 23 and 206. Same below for the full scans.

We have repeated the model training with re-balanced total numbers of cases/controls in response to Reviewer 1's comment on the imbalance. All new numbers have been reported, along with breakdowns of how many images contained clips and how many did not.

3) page 3 line 27: Verify that you split at the scan not slice level - it sounds it but need to verify.

We have updated the wording to reflect this verification.

4) tables 1-3: All of these need the data to be compared statistically for significant differences - likely a repeated measures ANOVA.

We may need further clarification on this suggestion. We have consulted two statisticians who do not think that repeated measures ANOVA is a suitable analysis in this application. We are very happy to discuss this point further to better understand the suggestion.

5) Delete the figures of ROC curves where data already in the tables.

We have deleted these figures.

Reviewer 1:

Introduction: I would recommend that the paragraph beginning pg 2, line 17 would be more suitable for the discussion section, rather than the introduction.

We have moved the paragraph as suggested.

Materials and Methods: Would recommend that the RIS manufacturer / model is identified for clarity to the reader.

We have now identified the RIS manufacturer/model.

Materials and Methods: it would be good to state in the results section if all images were agreed upon by 2 radiologists, or if and for how many a 3rd radiologist was needed to confirm.

We agree that this would be a useful inclusion – unfortunately we do not have access to this information.

Materials and Methods: On page 2, line 48 to page 3, line 2 the authors describe that 'for each scan with an aneurysm clip present, a scan with no aneurysm clip was identified' for matched controls. If this is the case, why were there 140 datasets with aneurysm clips and 122 without as identified in the abstract, rather than an equal number of each?

The explanation is that some control scans were lost due to issues such as duplication and corruption of files. However, in light of this comment, we have re-balanced the dataset and re-run the analysis. All reported counts and metrics now reflect this change. There are now 140 scans with aneurysm clips and 140 without, a precise balance which we feel enhances the quality of the analysis.

Materials and Methods: Spilt (page 3) it is not clear how this number was reduced from 262 to 229 localiser / 202 reconstructed CT scans. It would also be useful to the reader to understand the number of image datasets containing / not containing aneurysm clips for each in the methods section.

We have updated the wording to clarify the selection of subsets (sagittal localizers and scans reconstructed using a bone kernel). We have also included counts of those with aneurysm clips and those without.

Materials and Methods: How were the holdout test sets for both localisers and reconstructed datasets chosen?

We have updated the wording to reflect that images “were randomly divided at a scan level” into the training and holdout sets.

Materials and Methods: Image Preprocessing (Page 4). Please clarify how the images were cropped, i.e. manual / automatic, and with what software as the preprocessing performed.

We have updated the wording to reflect that “the images were preprocessed before model input by a deterministic automatic pipeline developed in Python using tools from OpenCV, SciPy and scikit-image.”

Results: Localiser results - recommend that the clip / no clip dataset numbers is described in methods rather than results. Again, it would be useful for the reader to understand why there were unequal numbers of clip / no clip datasets.

Total numbers have now been balanced as above, and clip/no clip numbers have been reported in methods.

Results: 3.1.1 Is useful in determining what caused the false positives, i.e. fillings. However, could datasets containing fillings not have been used in the training datasets within the 'no clip' dataset to mitigate this?

Before the decision was made to crop images to remove fillings, fillings were included in both the clip and no clip datasets. This resulted in false positives, which is why the decision was made to exclude the mandible.

Results: As with the localiser dataset, the reader would benefit from a table similar to Table 1, showing the performance of different base models. This is not discussed and would be useful to see how the models compare between localiser and reconstructed datasets and if different, why this could be.

We could not use a base model for the 3D networks as no such pre-trained models exist. The model was trained from scratch, which is why no similar analysis could be conducted.

Discussion: it would be good to understand if the model truly is a detector of aneurysm clips, or merely metallic implants that has only been trained on aneurysm clips? In which case, are these really false positives if the aim is to detect any metallic implant for the purposes of MR safety, which could then be determined by the appropriate specialist visually inspecting the images? This is alluded to within the discussion on page 12, line 39 and further evident in the heat maps within the supplementary material figure 2, where the skull flap fixing plates and skin clips also elicit hot spots on the probability maps. It would be interesting to see the results for patients who underwent neurosurgery and had skin clips / burrhole plate etc, without aneurysm clips.

This is a very helpful point, and we are particularly grateful for the thorough attention paid to the supplementary material. We have updated the wording to include the reviewer's astute observation regarding the skull fixing plates and skin clips. Having re-run the analysis on the newly balanced dataset, patients without aneurysm clips but with other implants are now represented in the supplementary material as false positives. We have further updated the wording to acknowledge that "future work could assess these models on a CT head dataset incorporating a wider range of metallic implants, to analyse whether models trained to detect aneurysm clips specifcally generalize to metal implant detection more broadly."

Discussion: The discussion about the use of ensemble methods is interesting, and would be good to know why the authors did not take this step, already having data for 5 models?

Reflecting on this observation, we have removed the reference to an ensemble of the five trained models. Our intention in reporting these "simple ensemble" results was to demonstrate the potential power of weak learners. However, upon consideration, we think that this confuses the suggestion we are making, which would be a much more extensive investigation of ensembling. We have updated the wording to reflect that this might include "different learning algorithms, as well as bagging or boosting approaches".

Conclusion: neglects to summarise the findings for the reconstructed CT scans. Recommend including this also.

Reconstructed CT scans now included in conclusion.

Figure 4(a) + (b): whilst the mean ROC curves are shown, the 5 models are indistinguishable. It would be useful to either label all model lines for clarity, or at the least, identify the chosen model curve please. Likewise with figure 7.

As per editor's suggestion, these figures have been removed.

Reviewer 2

Clearly, the clips are really small and it is expected that 3D networks would heavily overfit and not perform well, therefore, some techniques against that must be investigated.

We appreciate this suggestion, which would be an interesting exercise and a potential direction of future work. We think it is beyond the scope of this application, which found that the use of localizer images yielded better sensitivity results and was drastically less computationally expensive. The suggested refinement of the 3D technique would be valuable had the 2D technique not been so successful.

Is it the first work in this area? No comparisons to other methods are done.

This is the first work in this area. We have now clarified this at the start of the discussion, where we note the work done in detection of other implants, and acknowledge "the successful implementation of deep learning for implant detection is continued in this application, the first to use deep learning to detect aneurysm clips."

Reviewer 3

The authors mention the use of a GPU for training. Would be nice to specify which GPU was used.

We have now included details of the GPU.

The authors used five models, namely VGG16, Inception V3, Xception, DenseNet and MobileNet V2. While the results (at least for the localizer images) speaks for themselves, the authors should explain why the included those five models in their study.

We have updated the wording to explain that these are well-established models.

It would be interesting to compare the selected models not just for their performances, but also with respect to other parameters (e.g. model parameters, inference time, FLOPs etc.)

We have now analysed these parameters and included them in Table 1.

The selected network architectures are by today's standards a bit outdated. Why did they not use something more modern like UNet or its derivatives or even a transformer architecture?

We agree that the networks are not amongst the latest developments in ML-based image classification - however they often perform extremely well, as demonstrated in this application. We have now updated the discussion of future work to suggest that different learning algorithms could be considered in an ensemble: transformers are such an architecture that could be included.

The authors mention that they didn't use a pre-trained 3D-CNN, because they didn't find one readily available. Would it be an option to pre-train one by yourself or do you lack compute ressources to do it by yourself?

We have trained a 3D network from scratch to classify presence of aneurysm clips in this application. The pre-training of a network that could have much broader transfer learning use (along the lines of the 2D networks trained on ImageNet) would require vast computational and imaging resources which we do not possess.

They mention in the text that they used flipping as the only data augmentation technique and that additional augmentation didn't improve the performance. However, they conclude that "This suggests that there might be scope to improve the models' performance even further with access to more training data". Why does augmenting the data further does not improve the performance, yet more training data would?

It is true that we might not be able to improve the models' performance with access to more data (especially as the performance is already very high). Our reasoning was that a dataset which contained more examples of aneurysm clips would give the models more information than artificially increasing the size of the dataset using augmentation. To clarify, we have removed this wording from the discussion section on explainability, leaving a similar point in the limitations that "if it were possible to obtain more data this might enable the development of even more accurate models in training, and enable more representative assessment of models in the holdout set".

The authors should report the (hyper-)parameters used (e.g learning rate, batch size, probability of horizontal flip etc.)

These details have been added.

The size of the dataset and the lack of external validation are limitations that the authors also mention in their paper. Regarding external validation, they state that their "research team is in the process of planning and gaining governance clearance for such accessible studies". I think it would be better to defer the publication of this journal article to include the results of this study.

We appreciate and understand this suggestion. We think that the reported investigation is the initial stage of an important application which we hope to progress in a timely manner. Obtaining governance clearance for data access outside the routine care team can be an involved and

long-winded process and we have no guarantee of the timelines for this. This work will require applications for additional funding which will be enhanced by peer reviewed publications in respected journals. On balance we believe that expediting dissemination of these results is in the interests of both patients and the research community.

Deep learning detection of aneurysm clips for magnetic resonance imaging safety

Megan Courtman[1]*, Daniel Kim[2], Huub Wit[3], Hongrui Wang[4],

Lingfen Sun[1], Emmanuel Ifeachor[1], Stephen Mullin[5],

Mark Thurston[4]

[1] Faculty of Science and Engineering, University of Plymouth, Plymouth, PL4 8AA, United Kingdom

[2] Department of Radiology, Royal Cornwall Hospitals NHS Trust, Truro, TR1 3LJ, United Kingdom

[3] Department of Radiology, Torbay and South Devon NHS Trust, Torquay, TQ2 7AA, United Kingdom

[4] Department of Radiology, University Hospitals Plymouth NHS Trust, Plymouth, PL6 8DH, United Kingdom

[5] Plymouth Institute of Health and Care Research, University of Plymouth, Plymouth, PL4 8AA, United Kingdom

* Corresponding author. Email: megan.courtman@plymouth.ac.uk. ORCID: 0000-0002-8984-7798.

Statements & Declarations

Funding

This work was supported by the Engineering and Physical Sciences Research Council under Grant EP/T518153/1 and the National Institute for Health and Care Research under Grant PEN/006/005/A.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Megan Courtman: Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing

Daniel Kim: Data Curation, Investigation

Huub Wit: Data Curation, Investigation

Hongrui Wang: Data Curation, Software, Writing – Review & Editing

Lingfen Sun: Conceptualization, Supervision, Writing – Review & Editing

Emmanuel Ifeachor: Conceptualization, Supervision, Writing – Review & Editing, Funding Acquisition

Stephen Mullin: Conceptualization, Supervision, Writing – Review & Editing

Mark Thurston: Conceptualization, Methodology, Software, Data Curation, Investigation, Supervision, Writing – Review & Editing

Ethics approval

Ethical approval was granted on 15 October 2019 by HRA and Health and Care Research Wales.