Research Highlights

Using machine learning for rapid intracranial haemorrhage segmentation on axial computed tomography slices

Frederick Renyard
Year 3, Medicine, University of Bristol
Email: tw18869@bristol.ac.uk

Abstract

Introduction This study aimed to explore ways in which machine learning can be used for rapid segmentation and explainable classification of intracranial haemorrhage and discusses other potential implementations of the technology.

Methods An existing architecture was applied to a dataset of axial, brain-window slices of haemorrhagic and non-haemorrhagic scans, with radiologist masks over areas of diagnosed haemorrhage.

Results As a classifier, the model used in this study achieved an area under a precision-recall curve value of 0.93 (95% CI: 0.925, 0.935) and a maximum F1 score of 0.875 (95% CI: 0.817, 0.933) on the test dataset. When used for segmentation, the model achieved a maximum correlation coefficient of 0.80 ($p < 0.001$). When used to predict haemorrhage area, the intersection over union score was 0.64 (95% CI: 67.5, 75.7).

Conclusion The model used in this study quickly produces inferences, which is suited to real-time imaging modalities, such as ultrasound. However, more training data is required to improve the model, and external validation should be conducted to confirm the results.

Abbreviations

AUC - Area under the curve
CT - Computed tomography
IoU - Intersection over Union
ROC - Receiver-operator characteristic

Introduction

Intracranial haemorrhage is both a life-threatening and time-sensitive diagnosis, with one year mortality ranging between 51% and 65% and half of deaths occurring within two days.[1] Computed tomography (CT) scans are done routinely in trauma and stroke settings, in which time to diagnosis is crucial. The use of machine learning has been shown to reduce reporting time in trauma and stroke clinical settings, along with reducing length of stay in the Emergency Department.[2]

Machine learning is the use of statistical inference algorithms to predict diagnoses. The technology could provide an accessible method for rapidly extracting interpretations of data to improve patient outcomes.[2] The aim of this study was to develop an algorithm for imaging analysis via a method known as semantic segmentation, whereby an algorithm assigns a value to each part of an image according to how likely the part belongs to a class, such as "haemorrhage" or "fracture". These inferences are learnt from a dataset of pre-segmented images[3] (see **Figure 1**). This technique is already being used in biomedical sciences, from assigning cell types in microscopy to assisting brain mapping in neuroscience.[4,5]

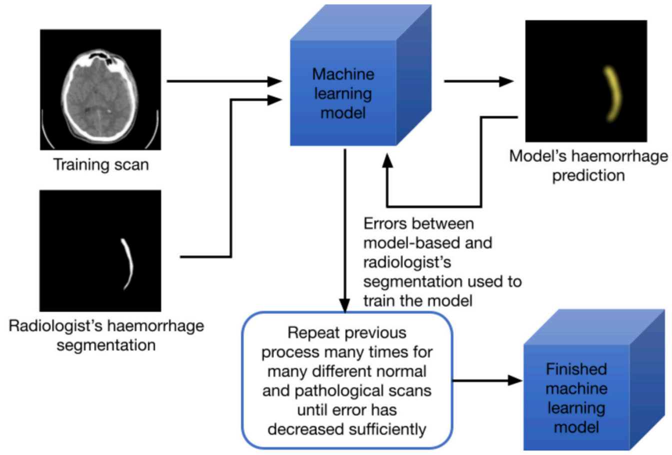

Training scan
Machine learning model
Model's haemorrhage prediction
Radiologist's haemorrhage segmentation
Errors between model-based and radiologist's segmentation used to train the model
Repeat previous process many times for many different normal and pathological scans until error has decreased sufficiently
Finished machine learning model

Figure 1. Machine-based learning. General overview of how a machine learning model for semantic segmentation is trained to produce a functional model. Scan images from Hssayeni (2019).[6]

Methods

A technical description of the results can be found in the GitHub repository: https://github.com/freddie-renyard/CT-Segmentation-UNet. The repository includes information on all code used in data pre-processing, model design, and post-training analysis, along with the final model and a full set of results as graphs.

The model was developed using Google's TensorFlow library. The structure of the model was based on an existing architecture known as U-Net, which was developed for use in biomedical applications due to its ability to learn from small datasets. The architecture is small, requiring less computational resources to run the model.[4]

The dataset was sourced from the data science website Kaggle (https://www.kaggle.com/vbookshelf/computed-tomography-ct-images), being made publicly available by Murtadha Hssayeni. The images were collected over a 7 month period at Al Hilla Teaching Hospital, Iraq, as part of a study by Hssayeni.[6] Ethical approval was granted for the study by the Iraqi Ministry of Health and all data was completely anonymised.[6]

The dataset contains the bone and brain windows of around 30 slices of axial CT scans of 82 patients, totalling 2500 images for each window. A description of patient demographics can be found in **Table 1**. The scans contain different types of intracranial haemorrhage, as well as non-pathological images (see **Table 2**). They also include segmentations of areas where there is intracranial haemorrhage present in each slide, which are annotated by radiologists.

Table 1. Patient demographics.

Number of patients	82
Mean age (±standard deviation)	27.84 ± 19.52
Maximum age	72 years
Minimum age	1.7 weeks
Male:female ratio	1.28:1

Table 2. Frequency of different haemorrhage types in the dataset.

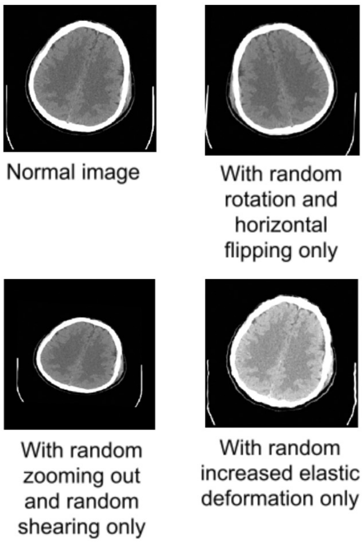
Haemorrhage type	Percentage of patients with diagnosis
No haemorrhage	56.1%
Intraventricular	6.1%
Intraparenchymal	19.5%
Subarachnoid	8.5%
Epidural	25.6%
Subdural	4.9%
Fractures present	26.8%

Before training the model, the data was pre-processed and sorted into pathological and non-pathological classes. The brain windows were used for model training. Since testing data was not given in the dataset, four full cases (amounting to 127 images) were withheld from the training dataset to serve as validation data and testing data, enabling the model to be evaluated with data that it had not been trained on.

In order to expand the small dataset, extensive data augmentation was used. These were all biologically plausible modifications.[7] The techniques used were:

1. Horizontal flipping: reverses left and right
2. Rotation: randomly rotates the image by up to 40 degrees, which frequently occurs due to suboptimal patient positioning during imaging
3. Shearing: randomly shears the image
4. Zooming: randomly zooms the image, simulating different patient sizes or scan setups
5. Brightness: randomly changes brightness, simulating different scan settings
6. Elastic deformation: randomly deforms the image, simulates patient tissue differences (note, this has been used to train more accurate models in radiological settings due to the biologically plausible method of deformation).[7]

See **Figure 2** for examples of scans before and after the techniques have been applied.


Normal image

With random rotation and horizontal flipping only

With random zooming out and random shearing only

With random increased elastic deformation only

Figure 2. An example of the data augmentation techniques used to augment the relatively small dataset used in this study. Note: the intensity of the elastic deformation presented was twice what was used during training. Scan image from Hssayeni (2019).[6]

The data in the dataset was unbalanced, consisting of 2,182 normal images against 318 pathological slides. The type of network used (convolutional neural network) has been shown to produce more accurate models when trained on a dataset that has been made to be balanced by duplicating images.[8] Thus, the dataset was balanced out via pathological slide duplication, bringing the number of pathological images to 2,024.

The model's architecture was created using code from an existing implementation by the GitHub user 'nikhilroxtomar';[9] the code for model training and data pre-processing was written separately. The model is trained by inputting image data along with annotations made by radiologists. Parameters of the network are modified until maximal performance metrics are reached. The model contains 1,962,337 modifiable parameters.

For training, the images underwent data augmentation, as detailed above, increasing the number of training images used per epoch to

5000. The validation dataset consisted of 127 images from 4 random cases and was used to evaluate the model at the end of each epoch of training data. To allow evaluation of the final model on the validation data, the model did not learn from this validation data. This process was repeated 144 times (144 epochs of training), with the images being used in batches of 16. Training took approximately 18 hours on an NVIDIA® GeForce® RTX 2070 SUPER®. The model was saved every 5 epochs to allow for the evaluation of successive models; this ensured that overfitting of the model to the data did not occur.[10] The error during training on both data partitions is shown in **Figure 3**. Unexpectedly, the error was greater for the training data than the validation data. This is likely to be attributed to the fact that extensive data augmentation had been applied to the training data, but no augmentation was applied to the validation data, as would be the case if the model were to be evaluated using clinical data.

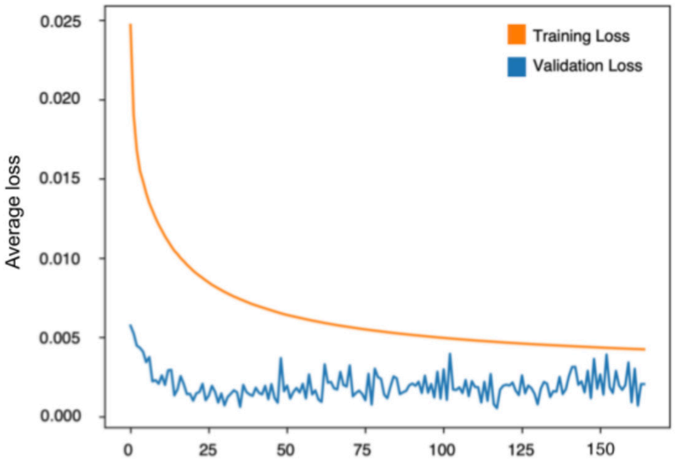


Figure 3. A graph of the mean loss of the model during each epoch (training loss) or with the validation data (validation loss).

Final evaluation of the model was performed on the four cases that were withheld from the original dataset (i.e. data that it had been trained on). The model was evaluated for its ability to classify scans as haemorrhagic and normal, and for its ability to segment the pixels into haemorrhagic and normal classes. The testing data included both normal scans and scans with haemorrhages.

Results

Classification The main metric used to evaluate the model's performance as a classifier was a precision-recall curve. The model's output was haemorrhage probabilities for each pixel. The mean of all predictions across each image in the test dataset was calculated and processed using a threshold to obtain a binary classification. Different precision and recall values were produced depending on where the threshold was set (see **Figure 4**). The area under the curve (AUC) represents the model's average precision, which for the classifier, was 0.93 (95% CI: 0.925, 0.935; p = 4.54x10-^{2}).

Another metric used to evaluate the classification model was the F1 score, which was derived from the precision and recall values. This score is frequently used in the evaluation of classifiers.[11] The highest F1 score achievable across all thresholds on the precision-recall curve was 0.875 (95% CI 0.817, 0.933; p < 0.001). At this score, the precision was 82.4% and the recall was 93.3%.

Segmentation

The ability of the area of the predictions to estimate haemorrhage size was analysed. The mean value of all the pixels in each original mask in the testing dataset was calculated, along with the mean value of all the pixels in each predicted mask. This was used as a marker of mask area. The predicted mask was given a threshold to make the output binary, as described above. The correlation between the pairs of mean values for each image was calculated over a range of classification thresholds, and the threshold with the highest Pearson correlation coefficient was chosen. This threshold was optimal for the validation data (note that further evaluation on clinical data would be needed to optimise this for different applications). As previously described, the model was evaluated on the validation dataset (n = 127), which was set aside from the training data. The correlation coefficient at this threshold was 0.80 (p <6.5x10-^{29}).

Having determined the optimal threshold for optimum prediction, the accuracy of segmentation via this model at this threshold was analysed. The Intersection over Union (IoU) metric was used to evaluate the model's segmentation performance; this quantifies the overlap between the radiologist's segmentation of an image and the model-based segmentation.[12] Across the pathological cases in the testing data, 71.6% (95% CI: 67.5%, 75.7%) of the segmentation of images matched the radiologist's original segmentation; this was true even when the model falsely predicted haemorrhage in a normal scan.

A collection of randomly selected predictions from the model at the optimal threshold, alongside the associated scan and radiologist segmentation is presented in **Figure 5**.

Discussion

With regards to use of the model as a classifier of disease (haemorrhage), the precision-recall AUC value observed in this study (0.93) was in line with a similar study, by Monteiro *et al.*, which obtained an AUC value of 0.89 (CI 95%: 0.86, 0.91).[13] Notably, the model used in the previous trial was evaluated on an external validation set, increasing its reliability.[13] Monteiro and colleagues also used receiver-operator characteristic (ROC) curves in their analysis, which have been shown to be comparable to precision-recall curves.[14] However, precision-recall curves are better for analysing rare findings due to their ability to focus on uncommon pathological results versus frequent normal results.[15]

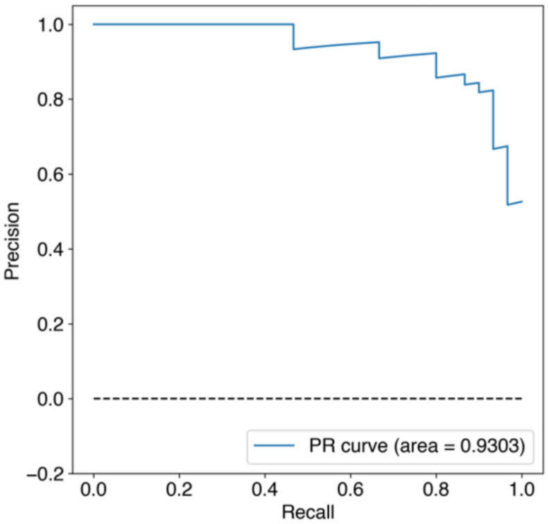


Figure 4. Precision-recall curve for haemorrhage classification. The precision-recall curve for the model as a scan-wise classification algorithm. PR, precision-recall.

In the present study, the model thresholds for both the F1 score and correlation coefficient were set at the maximum that could be achieved for the validation dataset. In order to assess the model's maximum thresholds in a clinical setting, a clinical dataset would be required for use in the model evaluation, using the same analysis as in this study to optimise the thresholds. This would set the balance between recall and precision needed for a given application. This technique is commonly used for this class of problems but it is difficult to optimise the model without large amounts of validation data.[16]

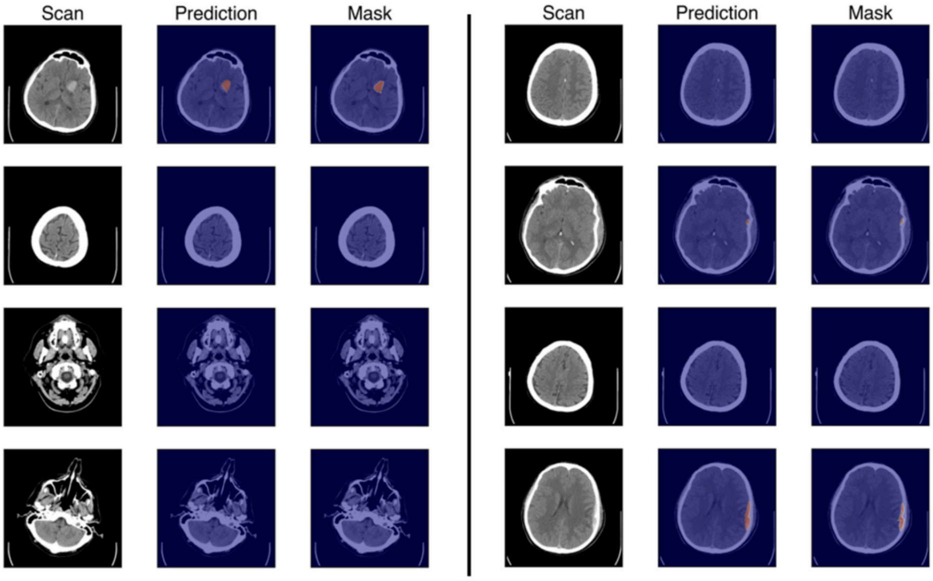


Figure 5. A collection of test scans, showing predictions from the final model, and radiologist segmentations. More predictions can be found in the GitHub repository: https://github.com/freddie-renyard/CT-Segmentation-UNet. Scan images from Hssayeni (2019).[6]

The maximum F1 score was 0.875; this value was derived from the same metrics as the precision-recall AUC but gives a less abstract indication of model performance. A precision of 82.4% at a recall value of 93.3% would restrict the model's use to highlighting areas of potential haemorrhage to a radiologist, rather than autonomous diagnosis. Validation with clinical data against radiological diagnosis would confirm or refute this claim. Another study, which focussed on classification of cranial CT scans, achieved lower precision and recall scores (67.8% and 61.0%, respectively), but the number of labels being classified was higher, at 9 categories rather than the 2 described here.[17]

The analysis of the model's ability to predict haemorrhage volume demonstrated the model to be moderately successful, with a significant, high positive correlation of 0.80.[18] However, this performance is inferior to more advanced algorithms; other models have achieved stronger correlations with testing data, with difference in haemorrhage volume estimations ranging from 0.07 ml to 2.09 ml for different haemorrhage classes.[13]

Upon use of the model in this study, segmentation performance on large haemorrhages (IoU = 72%) was lower compared with other medical algorithms; for example, when similar approaches were used to segment cervical muscles on ultrasound, IoU values of over 86% were obtained.[19] The UNet architecture has also been applied to abdominal CT data for segmentation of liver tumours, achieving scores of 92.6%.[20] Unfortunately, the model used in this study has not been tested on small haemorrhage volumes and, so, performance in this respect has not been determined.[13]

The main limitations of this study are:

1. The small dataset. Large datasets are needed for better neural network performance.[4] This has been demonstrated by Montiero *et al*.[13] who conducted a study that used a larger amount of data than this study for voxel-wise segmentation, which involves processing an entire scan with 3D data.
2. The lack of external validation. The only data that the model was analysed on is that of the dataset provided by Hssayeni *et al*.[6] Although the data analysed was withheld whilst training the model, evaluation of clinical data would be needed to determine the optimum threshold for use in the analysis.

Overall, the architecture used was not sufficiently complex to produce results at performance levels comparable to other studies that has used machine learning for analysis of CT data. However, its application to other real-time imaging modalities, like ultrasound, is more suitable as the model is small, potentially enabling predictions from the model (inferences) to be made locally and in real time using the computing resources available at hospitals. This has the benefit of local data processing, rather than sending data to a remote server, which could compromise data security.[21] In light of this, further experiments were performed to adapt the model for use on a portable device application (iPhone) using Apple's machine-learning framework. The model was able to infer segmentations of scans in around a fifth of a second, with potential real-time applications (see **Appendix 1** for details). The CT segmentation algorithm would be less useful in this format since clinicians are trained to pick up major CT abnormalities in emergency settings. However, if this model architecture was trained on ultrasound images and embedded into a portable device, it could assist with interpretation.

Conclusion Machine learning models are often criticised for their black box characteristics, producing diagnoses with no explanation.[22] Segmentation-based models provide an alternative way of analysing scans using machine learning. CT is a good imaging modality for use with an advanced model that can extract large quantities of analytical information from the scan, despite being slow to execute, as the scan is performed once and the data can be saved to be viewed later. Ultrasound would be a better imaging modality for implementation of this model's architecture (U-Net) as the images change in real-time, and this model is small enough to produce rapid, offline inferences (so-called 'AI at the edge').[23] This could provide a heatmap of the image to help identify structures for ultrasound-guided nerve blocks or IV access, helping clinicians to interpret difficult imaging.

In general, convolutional neural networks will perform better when more data is available.[8] Developing better algorithms for machine learning models majorly depends on the availability of large datasets of anonymised patient data. This is important in healthcare, where data is scarce due to confidentiality but where models must be trained to high levels of accuracy to ensure that diagnoses are not missed.

Acknowledgements Many thanks to Ben Lippold (University of Bristol, Bristol, UK) for giving the author access to his graphics card and TensorFlow environment to train the final model.

Contribution statement The author confirms that they were substantially involved in the analysis and interpretation of the data, along with being substantially involved in the model development. The author also drafted the work and revised it.

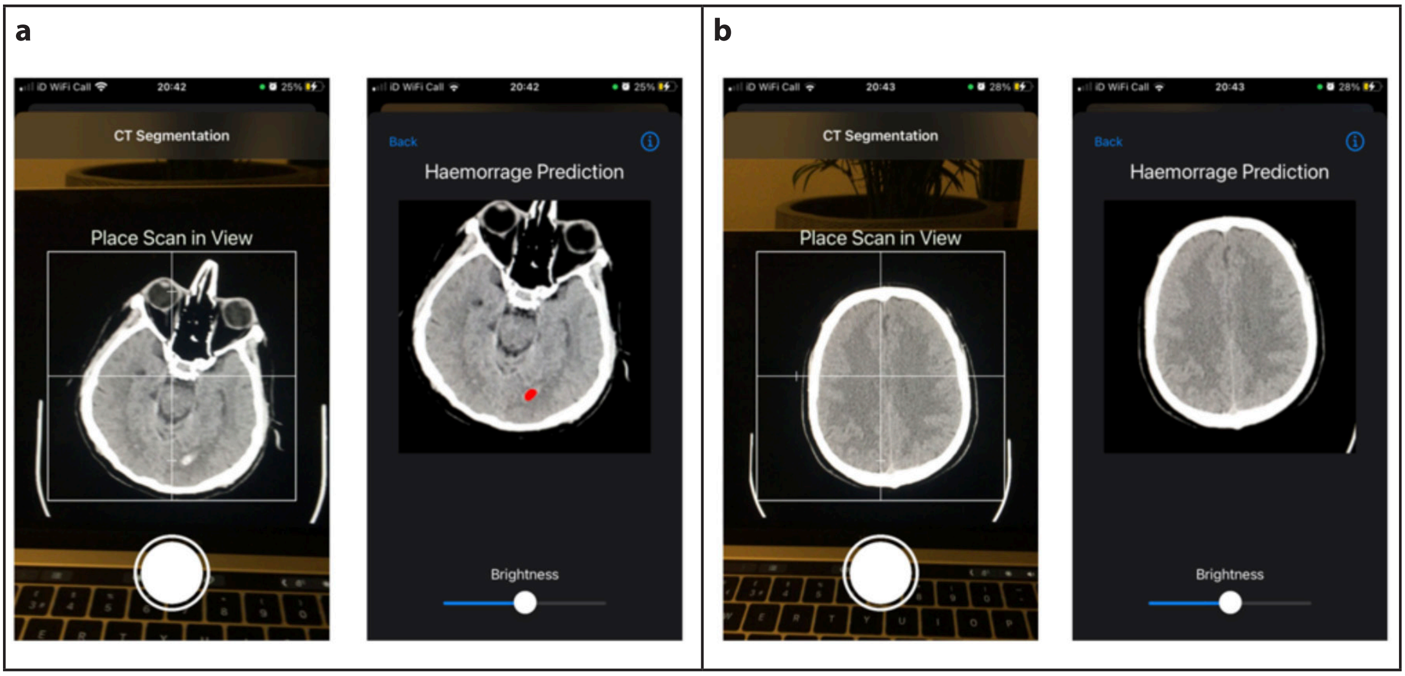
and reviewed by students, and the Editorial Board is composed of students. Thus, this journal has been created for educational purposes and all content is available for reuse by the authors in other formats, including peer-reviewed journals.

References

1. Rymer MM. Hemorrhagic stroke: intracerebral hemorrhage. Mo Med. 2011;108(1):50-54.
2. National Institute for Health and Care Excellence (2020). Artificial intelligence for analysing CT brain scans. Available from: https://www.nice.org.uk/advice/mib207/resources/artificial-intelligence-for-analysing-ct-brain-scans-pdf-2285965396121029. Accessed: 8 April 2021.
3. Ulku I, Akagunduz E (2020). A survey on deep learning-based architectures for semantic segmentation on 2D images. Available from: https://arxiv.org/abs/1912.10230. Accessed: 8 April 2021.
4. Ronnenberger O, Fischer P, Brox T (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Available from: https://arxiv.org/abs/1505.04597v1. Accessed: 8 April 2021.
5. Scheffer L, Xu C, Januszewski M, et al. A connectome and analysis of the adult Drosophila central brain. eLife 2020;9:e57443
6. Hssayeni, M (2019). Computed Tomography Images for Intracranial Hemorrhage Detection and Segmentation (version 1.0.0). Available from: https://doi.org/10.13026/w8q8-ky94. Accessed: 8 April 2021.
7. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation in breast cancer mass detection. Available from: https://ieeexplore.ieee.org/document/8333411. Accessed: 8 April 2021.
8. Hensman P, Masko D (2015). The Impact of Imbalanced Training Data for Convolutional Neural Networks. Available from: https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf. Accessed: 8 April 2021.
9. Tomar N (2020). U-Net Segmentation in Keras TensorFlow. Available from: https://github.com/nikhilroxtomar/UNet-Segmentation-in-Keras-TensorFlow. Accessed: 9 April 2021.
10. Tripathi M (2020). Underfitting and Overfitting in Machine Learning. Available from: https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning. Accessed: 8 April 2021.
11. Lipton ZC, Elkan C, Naryanaswamy B (2014). Optimal Thresholding of Classifiers to Maximize F1 Measure. In: Calders T, Esposito F, Hüllermeier E, et al (eds) Machine Learning and Knowledge Discovery in Databases.: ECML PKDD 2014. Lecture Notes in Computer Science, vol 8725. Springer, Berlin, pp 225-239.
12. Rezatofighi H, Tsoi N, Gwak J, et al (2019). Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. Available from: https://arxiv.org/abs/1902.09630. Accessed: 8 April 2021.
13. Monteiro M, Newcombe V, Francois M, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning : an algorithm development and multicentre validation study. The Lancet Digital Health. 2020;2(6):e314-22.
14. Davis J, Goadrich (2006). The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine Learning, 2006. Association of Computing Machinery, New York (NY), pp 233-240.
15. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
16. Google Machine Learning Crash Course (2020). Classfication: Thresholding. Available from: https://developers.google.com/machine-learning/crash-course/classification/thresholding. Accessed: 9 July 2021.
17. Li J, Fu G, Chen Y, et al. A multi-label classification model for full slice brain computerised tomography image. BMC Bioinformatics 2020;21(Suppl 6):200.
18. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal. 2012;24(3):69-71.
19. Cunningham RJ, Harding PJ, Loram ID. Real-Time Ultrasound Segmentation, Analysis and Visualisation of Deep Cervical Muscle Structure. IEEE Trans Med Imaging. 2017;36(2):653-665.
20. Jin Q, Meng Z, Sun C, et al. RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. Frontiers in Bioengineering and Biotechnology. 2020;8:605132
21. Naseem S (2020). Patient Bayesian Inference: Cloud-Based Healthcare Data Analysis Using Constraint-Based Adaptive Boost Algorithm. In: Niansheng Tang (eds) Bayesian Inference on Complicated Data. IntechOpen, London, pp 79-88.
22. Rudin C, Radin J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review. 2019;1(2). https://doi.org/10.1162/99608f92.5a8a3a3d.
23. Merenda M, Porcaro C, Iero D. Edge Machine Learning for AI-Enabled IoT Devices: A Review. Sensors 2020;20(9):2533.
24. Simonite T (2017). Apple's 'Neural Engine' Infuses the iPhone with AI Smarts. Available from: https://www.wired.com/story/apples-neural-engine-infuses-the-iphone-with-ai-smarts/. Accessed: 8 April 2021.

Appendix 1: Adapting the model for use on an iPhone using Apple's machine-learning framework

In order to test the model's speed on a device with a relatively small amount of computational power, a model was created for use on the iPhone using Apple's Core Machine Learning tools. This is a feature that the company is gradually making more powerful, along with adding more hardware to speed up AI applications on the device.[24] Some screenshots of the app's results on normal and haemorrhagic scans from the test data are given in **Supplementary Figure 1**.



Supplementary Figure 1. A collection of screenshots from the test application built on the iPhone. The scans are derived from the test dataset used in the main study (https://www.kaggle.com/vbookshelf/computed-tomography-ct-images). (**a**) A scan that is positive for haemorrhage, as confirmed by a radiologist. (**b**) A normal scan.

The app was tested on the iPhone 6S and, even with this older hardware, the model was able to process a request and display the result in a mean time of 0.218 seconds (standard deviation = 0.017 seconds; n = 5). This indicates that the model could theoretically run at around 4 and a half frames per second on the iPhone 6S, which could be improved with use of newer devices with specific AI hardware. These sorts of applications of AI on the edge could be used for real-time interpretation of imaging mediums, like ultrasound, which could provide an added layer of understanding for images that are often difficult to interpret.