Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head CT Scans

Sasank Chilamkurthy[1], Rohit Ghosh[1], Swetha Tanamala[1], Mustafa Biviji[2], Norbert G. Campeau[3], Vasantha Kumar Venugopal[4], Vidur Mahajan[4], Pooja Rao[1], and Prashant Warier[1]

[1]Qure.ai, Mumbai, IN
[2]CT & MRI Center, Nagpur, IN
[3]Department of Radiology, Mayo Clinic, Rochester, MN
[4]Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, IN

Abstract

Importance Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms.

Objective To develop and validate a set of deep learning algorithms for automated detection of following key findings from non-contrast head CT scans: intracranial hemorrhage (ICH) and its types, intraparenchymal (IPH), intraventricular (IVH), subdural (SDH), extradural (EDH) and subarachnoid (SAH) hemorrhages, calvarial fractures, midline shift and mass effect.

Design and Settings We retrospectively collected a dataset containing 313,318 head CT scans along with their clinical reports from various centers. A part of this dataset (Qure25k dataset) was used to validate and the rest to develop algorithms. Additionally, a dataset (CQ500 dataset) was collected from different centers in two batches B1 & B2 to clinically validate the algorithms.

Main Outcomes and Measures Original clinical radiology report and consensus of three independent radiologists were considered as gold standard for Qure25k and CQ500 datasets respectively. Area under receiver operating characteristics curve (AUC) for each finding was primarily used to evaluate the algorithms.

Results Qure25k dataset contained 21,095 scans (mean age 43.31; 42.87% female) while batches B1 and B2 of CQ500 dataset consisted of 214 (mean age 43.40; 43.92% female) and 277 (mean age 51.70; 30.31% female) scans respectively. On Qure25k dataset, the algorithms achieved an AUC of 0.9194, 0.8977, 0.9559, 0.9161, 0.9288 and 0.9044, for detecting ICH, IPH, IVH, SDH, EDH and SAH respectively. AUCs for the same on CQ500 dataset were 0.9419, 0.9544, 0.9310, 0.9521, 0.9731 and 0.9574 respectively. For detecting calvarial fractures, midline shift and mass effect, AUCs on CQ500 dataset were 0.9244, 0.9276 and 0.8583 respectively, while AUCs on Qure25k dataset were 0.9624, 0.9697 and 0.9216 respectively.

Conclusions and Relevance This study demonstrates that deep learning algorithms can accurately identify head CT scan abnormalities requiring urgent attention. This opens up the possibility to use these algorithms to automate the triage process. They may also provide a lower bound for quality and consistency of radiological interpretation.

1 Introduction

Non-contrast head CT scans are among the most commonly used emergency room diagnostic tools for patients with head injury or in those with symptoms suggesting a stroke or rise in intracranial

pressure. Their wide availability and relatively low acquisition time makes them a commonly used first-line diagnostic modality [1, 2]. The percentage of annual US emergency room visits that involve a CT scan has been increasing for the last few decades [3] and the use of head CT to exclude the need for neurosurgical intervention is on the rise[4].

The most critical, time-sensitive abnormalities that can be readily detected on CT scan include intracranial hemorrhages, raised intracranial pressure and cranial fractures. A key evaluation goal in patients with stroke is excluding an intracranial hemorrhage. This depends on CT imaging and its swift interpretation. Similarly, immediate CT scan interpretation is crucial in patients with a suspected acute intracranial hemorrhage to evaluate the need for neurosurgical treatment. Cranial fractures, if open or depressed will usually require urgent neurosurgical intervention. Cranial fractures are also the most commonly missed major abnormality on head CT scans [5, 6], especially if coursing in an axial plane.

While these abnormalities are found only on a small fraction of CT scans, streamlining the head CT scan interpretation workflow by automating the initial screening and triage process, would significantly decrease the time to diagnosis and expedite treatment. This would in turn decrease morbidity and mortality consequent to stroke and head injury. An automated head CT scan screening and triage system would be valuable for queue management in a busy trauma care setting, or to facilitate decision-making in remote locations without an immediate radiologist availability.

The past year has seen a number of advances in application of deep learning [7, 8, 9, 10, 11, 12] for medical imaging interpretation tasks, with robust evidence that deep learning can perform specific medical imaging tasks including identifying and grading diabetic retinopathy [13] and classifying skin lesions as benign or malignant [14] with accuracy equivalent to specialist physicians. Deep learning algorithms have also been trained to detect abnormalities on radiological images such as chest radiographs[8, 9], chest CT [15, 16] and head CT [10, 11] through 'classification' algorithms; as well as to localize and quantify disease patterns or anatomical volumes [17, 18, 19] through 'segmentation' algorithms.

The development of an accurate deep learning algorithm for radiology requires – in addition to appropriate model architectures – a large number of accurately labeled scans that will be used to train the algorithm[20]. The chances that the algorithm generalizes well to new settings increase when the training dataset is large and includes scans from diverse sources[21, 22] .

In this manuscript, we describe the development, validation and clinical testing of fully automated deep learning algorithms that are trained to detect abnormalities requiring urgent attention from non-contrast head CT scans. The trained algorithms detect five kinds of intracranial hemorrhages (ICH) namely intraparenchymal (IPH), intraventricular (IVH), subdural (SDH), extradural (EDH) and subarachnoid (SAH), and calvarial/cranial vault fractures. Algorithms also detect mass effect and midline shift, both used as indicators of severity of the brain injury.

2 Methods

2.1 Datasets

We retrospectively collected 313,318 anonymous head CT scans from several centers in India. These centers, which included both in-hospital and outpatient radiology centers, employ a variety of CT scanner models (shown in Table 1) with slices per rotation ranging from 2 to 128. Each of the scans had a clinical report associated with them which we used as the gold standard during the algorithm development process.

Of these scans, we earmarked scans of 23,163 randomly selected patients (Qure25k dataset) for validation and used the scans of rest of the patients (development dataset) to train/develop the algorithms. We removed post-operative scans and scans of patients less than 7 years old from the Qure25k dataset. This dataset was *not* used during the algorithm development process.

A clinical validation dataset (referred as CQ500 dataset) was provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics (CARING), New Delhi, India. This dataset was a subset of head CT scans taken at various radiology centers in New Delhi. Approximately, half of the centers are stand-alone outpatient centers and the other half are radiology departments embedded in large hospitals. There was no overlap between these centers and the centers from where

Dataset	CT Scanner Models
Qure25k & Development	GE BrightSpeed, GE Brivo CT315, GE Brivo CT385, GE HiSpeed, GE LightSpeed, GE ProSpeed, GE Revolution ACTs, Philips Brilliance, Siemens Definition, Siemens Emotion, Siemens Sensation, Siemens SOMATOM, Siemens Spirit
CQ500	GE BrightSpeed, GE Discovery CT750 HD, GE LightSpeed, GE Optima CT660, Philips MX 16-slice, Philips Access-32 CT

Table 1: Models of CT scanners used for each dataset

the development dataset was obtained. CT scanners used at these centers had slices per rotation varying from 16 to 128. Models of the CT scanners are listed in Table 1. The data was pulled from local PACS servers and anonymized in compliance with internally defined HIPAA[23] guidelines.

Similar to the development and Qure25k datasets, clinical radiology reports associated with the scans in the CQ500 dataset were available. Although, we do not use them as gold standards in this study, we use them for the dataset selection as described below.

We collected the CQ500 dataset in two batches (B1 & B2). Batch B1 was collected by selecting all the head CT scans taken at the above centers for 30 days starting from 20 November 2017. Batch B2 was selected from the rest of the scans in the following way:

1. A natural language processing (NLP) algorithm was used to detect IPH, SDH, EDH, SAH, IVH, calvarial fractures from clinical radiology reports.
2. Reports were then randomly selected so that there are around 80 scans with each of IPH, SDH, EDH, SAH and calvarial fractures.

Each of the selected scans were then screened for the following exclusion criteria:

- Patient should not have any post-operative defect such as burr hole/shunt/clips etc.
- There should be at least one non-contrast CT series with axial cuts and soft reconstruction kernel which covers the complete brain.
- Patient should not be less than 7 years old. Wherever age information is not available, it is roughly estimated from ossification degree of cranial sutures[24].

Follow up scans for a patient were *not* excluded in the selection process.

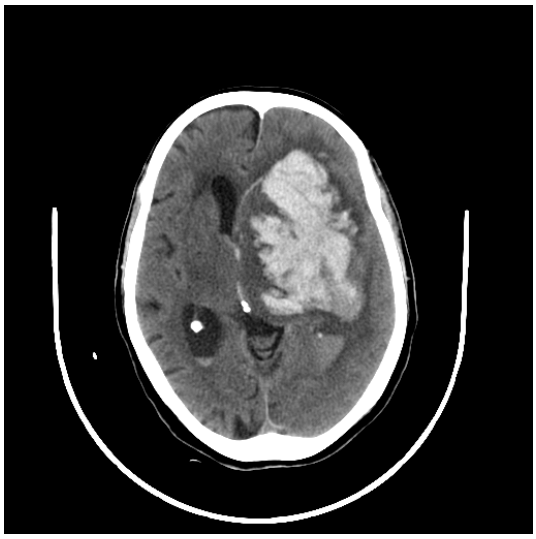
2.2 Reading the scans

Three senior radiologists[1] served as independent readers for the CT scans in the CQ500 dataset. They had corresponding experience of 8, 12 and 20 years in cranial CT interpretation. None of the 3 readers was involved in the clinical care or evaluation of the enrolled patients, nor did they have access to clinical history of any of the patients. Each of the radiologists independently evaluated the scans in the CQ500 dataset with the instructions for recording the findings and query resolution as per the supplement. The order of presentation of the scans was randomized so as to minimize the recall of the patients' follow up scans.

CT scans were reviewed on a custom web based viewer built upon orthanc[25] and OHIF framework[26]. Each reviewer could change the window level settings of the scans and had access to all the series present in a given CT scan.

Each of the readers recorded the following findings for each scan:

[1]Dr. Campeau was one of the three readers.



(a) Slice from a head CT scan

☑ Intracranial hemorrhage
- ☑ Intraparenchymal
- ☑ Intraventricular
- ☐ Subdural
- ☐ Extradural
- ☐ Subarachnoid

Location: ☑Left ☑Right

☐ Chronic

☑ Midline Shift

☑ Mass Effect

☐ Fracture
- ☐ Calvarial fracture

Remarks:

(b) Form used to record the findings

Figure 1: Schematic of the reading process for the CQ500 dataset

- The presence or absence of a intracranial hemorrhage and if present, its type(s) (intraparenchymal, intraventricular, extradural, subdural and subarachnoid), age (chronic or not) and cerebral hemisphere(s) affected (left, right).
- The presence or absence of midline shift and mass effect.
- The presence or absence of fractures. If present, if it is (partly) a calvarial fracture.

A text box was also provided to record any remark/observation which did not fit this decision flow. Schematic of the reading process along with the form used to record the findings is shown in Figure 1

Intra-axial presence of blood due to any etiology such as hemorrhagic contusion, tumour/infarct with hemorrhagic component was also included in the definition of intraparenchymal hemorrhage. Chronic hemorrhages were considered positive in this study. Mass effect was defined as any of the following: local mass effect, ventricular effacement, midline shift and herniation. Midline shift was considered positive if the amount of shift was greater than 5mm. If there is at least one fracture that extends into the skullcap, the scan is considered to have a calvarial fracture.

If unanimous agreement for each of the findings was not achieved by the three readers, the interpretation of majority of the readers was used as the final diagnosis.

On the development and Qure25k datasets, we considered clinical reports written by radiologists as the gold standard. However, these were written in free text rather than in a structured format. Therefore, a rule based natural language processing (NLP) algorithm was applied on radiologist's clinical reports to automatically infer the findings recorded above. We validated this algorithm on a subset of reports from the Qure25k dataset to ensure that the inferred information was accurate and could be used as gold standard.

2.3 Developing the algorithms

Deep learning is a form of machine learning where the model used is a neural network with a large number of (usually convolutional) layers[27]. *Training* this model requires large amount of data for which the ground truth is already known. Training is usually performed by an algorithm called back propagation[28, 29]. In this algorithm, model is iteratively modified to minimize the error between predictions of the model and the known ground truth for each data point.

One of the main challenges we faced in the development of the algorithms was the three dimensional (3D) nature of the CT scans. This was primarily due to an issue termed as 'curse of dimensionality' where the data required to train a machine learning algorithm scales exponentially with the

dimensionality of data [30]. Deep learning techniques have been extensively researched for the tasks of segmentation[31, 32, 33] and classification[34, 35] of two dimensional images. While the segmentation of 3D images is studied in multiple contexts [36, 37, 19], their classification is not as well investigated. One closely related problem is recognition of human actions from short video clips (because videos are three-dimensional with the time as the third dimension). Despite that this problem is well explored in the literature [38, 39, 40], there was no emergence of a front running architecture for this task[38]. Our approach of classification is closely related to that of Simonyan and Zisserman [39] and involved slice level and pixel level annotation of a large number of scans.

In this study, we trained separate deep learning models for each of the subtasks viz. intracranial bleeds, midline shift/mass effect and calvarial fractures which we describe below.

2.3.1 Intracranial Hemorrhage

Development dataset was searched for non-contrast head CT scans which were reported with any of the IPH, IVH, SDH, EDH, SAH and those with neither of these. Each slice in these scans was manually labeled with the hemorrhages that are visible in that slice. In all, 4304 scans (165809 slices) were annotated, of which number of scans (slices) with IPH, IVH, SDH, EDH, SAH and neither of these were 1787 (12857), 299 (3147), 938 (11709), 623 (5424), 888 (6861) and 944 (133897) respectively.

We used ResNet18[35], a popular convolutional neural network architecture with a slight modification to predict softmax based confidences[41] for the presence of each type of hemorrhage in a slice. We modified the architecture by using five parallel fully connected (FC) layers in place of a single FC layer. This design was based on the assumption that the image features[42] for detecting hemorrhages would be similar for all the hemorrhage types. The confidences at the slice-level are combined using a random forest[43] to predict the scan-level confidence for the presence of intracranial hemorrhage and its types.

We further trained a model to localize the following type of hemorrhages: IPH, SDH, EDH. Localization requires dense prediction[32] of presence or absence of bleed for every pixel in the scan. To train models for dense predictions, pixels corresponding to the each bleed were annotated for a subset of the above slice-annotated images to provide the ground truth for the model. This set contained 1706 images of which number of images with IPH, SDH, EDH and neither of these are 506, 243, 277 and 750 respectively. We used a UNet[33] based architecture for segmentation of each type of hemorrhage.

2.3.2 Midline Shift and Mass Effect

The algorithm for detecting midline shift and mass effect was very similar to the one for detecting intracranial hemorrhage. Each slice from select scans was labeled for the presence or absence of midline shift and mass effect in that slice. Overall, 699 scans (26135 slices) were annotated, of which number of scans (slices) with mass effect were 320 (3143) and midline shift were 249 (2074).

We used modified ResNet18 with two parallel fully connected layers to predict slice wise confidences for the presence of mass effect and midline shift respectively. These slice level confidences were thereby combined using a random forest to predict scan-level confidences for both the abnormalities.

2.3.3 Calvarial Fractures

Development dataset was searched for scans with calvarial fractures. Each slice in these scans was annotated by marking a tight bounding box around fractures. Number of scans annotated were 1119 (42942 slices) of which 9938 slices showed a calvarial fracture.

Slices along with target bounding box was fed into a DeepLab[31] based architecture to predict pixel-wise heatmap for fractures. Skull fractures are extremely sparse in this representation. Gradient flow in the backpropagation algorithm tends to be hindered for such sparse signals. We therefore employed hard negative mining loss [44, 45] to counter the sparsity of the annotation.

We engineered features representative of local fracture lesions and their volumes from the generated heatmaps of the whole scan. We then used these features to train a random forest to predict scan-wise confidence of presence of a calvarial fracture.

2.3.4 Preprocessing

For a given CT scan, we selected the non-contrast axial series which uses soft reconstruction kernel and resampled it so that slice thickness is around 5mm. We then resized all the slices of this series to a size of 224×224 pixels before passing to our deep learning models. Instead of passing the whole dynamic range of CT densities as a single channel, we windowed the densities by using three separate windows and stacking them as channels. Windows used were brain window ($l = 40, w = 80$), bone window ($l = 500, w = 3000$) and subdural window ($l = 175, w = 50$). This was because fracture visible in the bone window could indicate existence of an extra axial bleed in the brain window and conversely, presence of scalp hematoma in the the brain window could correlate with a fracture. Subdural window helps differentiate between the skull and an extra axial bleed that might have been indistinguishable in a normal brain window[46].

2.4 Evaluating the algorithms

The combined algorithms when run on a scan produces a list of 9 real valued confidence scores in the range of [0, 1] indicating the presence of the following findings: Intracranial hemorrhage and each of the 5 types of hemorrhages, midline shift, mass effect and calvarial fracture. As described before, the corresponding gold standards were obtained using majority voting for CQ500 dataset and by NLP algorithm of reports for Qure25k dataset.

For both CQ500 and Qure25k datasets, receiver operating characteristic (ROC) curves[47] were obtained for each of the above by varying the threshold and plotting true positive rate (i.e sensitivity) and false positive rate (i.e 1 - specificity) at each threshold. Two operating points were chosen on the ROC curve so that sensitivity ≈ 0.9 (high sensitivity point) and specificity ≈ 0.9 (high specificity point) respectively. Areas under the ROC curve (AUCs) and sensitivities & specificities at these two operating points were used to evaluate the algorithms.

2.5 Statistical Analysis

Sample sizes for proportions and AUCs were calculated using normal approximation and the method outlined by Hanley and McNeil [47] respectively. The prevalence of our target abnormalities in a randomly selected sample of CT scans tend to be low. This means that establishing the algorithms' sensitivity with a reasonably high confidence on an un-enriched dataset would require very large sample sizes. For example, to establish a sensitivity with an expected value of 0.7 within a 95% confidence interval of half length of 0.10, number of positive scans to be read ≈ 80. Similarly, for a finding with prevalence rate of 1%, to establish AUC within a 95% confidence interval of half length of 0.05, number of scans to be read ≈ 20000.

The Qure25k dataset used in this study was randomly sampled from the population distribution and had number of scans > 20000 following the above sample size calculations. However, constraints on the radiologist time necessitated the selective sampling strategy outlined in the section 2.1 for the CQ500 dataset. Manual curation of scans (by referring to the scans themselves) would have had selection bias towards more significant positive scans. We mitigated this issue by random selection where positive scans were determined from the clinical reports.

We generated confusion matrices for each finding at the selected operating points. We then calculated the 95% confidence intervals for sensitivity and specificity from these matrices using 'exact' Clopper-Pearson method[48] based on Beta distribution. Confidence intervals of area under the ROCs were calculated following the 'distribution-based' approach described by Hanley and McNeil [47]. On the CQ500 dataset, we measured the concordance between paired readers on each finding using percentage of agreement and the Cohen's kappa(κ) statistic[49]. In addition, we measured concordance between all the three readers on each finding using Fleiss' kappa(κ) [50] statistic.

Sample sizes, sample statistics and confidence intervals were calculated using scikit-learn[51] and statsmodels[52] python packages.

3 Results

Patient demographics and prevalences for each finding are summarized in Table 2. Qure25k dataset contained 21095 scans of which number of scans reported positive for intracranial hemorrhage and

Characteristic	Qure25k dataset	CQ500 dataset batch B1	CQ500 dataset batch B2
No. of scans	21095	214	277
No. of readers per scan	1	3	3
PATIENT DEMOGRAPHICS			
Age			
No. of scans for which age was known	21095	189	251
Mean	43.31	43.40	51.70
Standard deviation	22.39	22.43	20.31
Range	$7-99$	$7-95$	$10-95$
No. of females / No. of scans for which sex was known (percentage)	$9030/21064$ (42.87%)	$94/214$ (43.92%)	$84/277$ (30.31%)
PREVALENCE			
No. of scans (percentage) with			
Intracranial hemorrhage	2494 (11.82)	35 (16.36)	170 (61.37)
Intraparenchymal	2013 (9.54)	29 (13.55)	105 (37.91)
Intraventricular	436 (2.07)	7 (3.27)	21 (7.58)
Subdural	554 (2.63)	9 (4.21)	44 (15.88)
Extradural	290 (1.37)	2 (0.93)	11 (3.97)
Subarachnoid	611 (2.90)	9 (4.21)	51 (18.41)
Fractures	1653 (7.84)	8 (3.74)	31 (11.19)
Calvarial Fractures	992 (4.70)	6 (2.80)	28 (10.11)
Midline Shift	666 (3.16)	18 (8.41)	47 (16.97)
Mass effect	1517 (7.19)	28 (13.08)	99 (35.74)

Table 2: Dataset characteristics for CQ500 and Qure25k datasets.

calvarial fracture are 2494 and 992 respectively. CQ500 dataset included 491 scans of which batch B1 had 214 scans and batch B2 had 277 scans. B1 contained 35 and 6 scans reported with intracranial hemorrhage and calvarial fracture respectively. The same for B2 were 170 and 28 respectively.

Number of clinical reports analyzed in the selection process of the CQ500 dataset was 4462. Of these, number of selected scans for batches B1 and B2 were 285 and 440 respectively. Number of exclusions were 71 and 163 respectively resulting in a total of 491 scans. Reasons for exclusion were non availability of images (113), post operative scans (67), scan had no non-contrast axial series (32) and patient was less than 7 years old (22). Schematic of the dataset selection process of the CQ500 dataset is presented in Figure 2.

3.1 Qure25k dataset

Natural language processing (NLP) algorithm used to infer the findings from clinical reports in the Qure25k dataset was evaluated on a total of 1779 reports. Sensitivity and specificity of the NLP algorithm are fairly high; least performing finding was subdural hemorrhage with sensitivity of 0.9318 (95% CI 0.8134-0.9857) and specificity of 0.9965 (95% CI 0.9925-0.9987) while fracture was inferred perfectly with sensitivity of 1 (95% CI 0.9745-1.000) and specificity of 1 (95% CI 0.9977-1.000). Sensitivity and specificity for all the target findings on the evaluated 1779 reports is shown in Table 3a.

Table 4a and Figure 3 summarize the performance of deep learning algorithms on the Qure25k set. Algorithm achieved AUCs of 0.9194 (95% CI 0.9119-0.9269) on intracranial hemorrhage, 0.9244 (95% CI 0.9130-0.9359) on calvarial fracture and 0.9276 (95% CI 0.9139-0.9413) on midline shift. Algorithms performed the best on intraventricular hemorrhage with an AUC of 0.9559 (95% CI 0.9424-0.9694) and the worst on mass effect with an AUC of 0.8583 (95% CI 0.8462-0.8703).

At the high sensitivity operating point, sensitivities of the algorithms for intracranial hemorrhage, calvarial fracture and midline shift were 0.8990 (95% CI 0.8865-0.9105), 0.8982 (95% CI 0.8592-

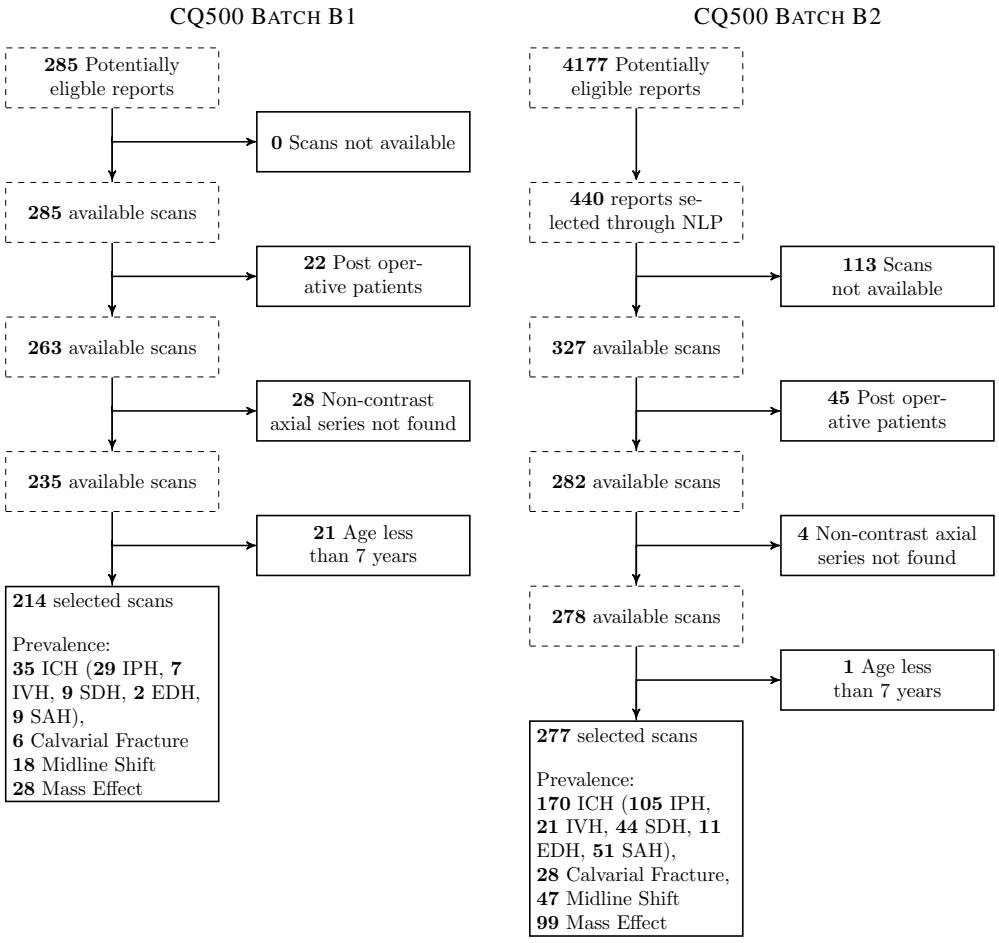


Figure 2: Dataset section process for batches B1 and B2 of the CQ500 dataset.

0.9197) and 0.8979 (95% CI 0.8724-0.9198) respectively. At this operating point, specificities for these findings were 0.7362 (95% CI 0.7298-0.7425), 0.7797 (95% CI 0.7739-0.7854) and 0.8698 (95% CI 0.8652-0.8744) respectively.

3.2 CQ500 dataset

Concordance between the three readers on the CQ500 dataset was observed to be the highest for intracranial hemorrhage (Fleiss' $\kappa = 0.7827$) and intraparenchymal hemorrhage (Fleiss' $\kappa = 0.7746$). Calvarial fracture and subdural hemorrhage had the lowest concordance with Fleiss' κ of 0.4507 and 0.5418 respectively. For each of the target findings, percentage agreement and Cohen's kappa between a pair of readers and Fleiss' kappa for all the readers is shown in 3b

The algorithms generally performed better on the CQ500 dataset than on the Qure25k dataset. AUCs, sensitivities and specificities are shown in Table 4b and ROCs are shown in 3. AUC for intracranial hemorrhage was 0.9419 (95% CI 0.9187-0.9651), for calvarial fracture was 0.9624 (95% CI 0.9204-1.0000), and for midline shift was 0.9697 (95% CI 0.9403-0.9991). Best AUCs were recorded on midline shift (0.9697 , 95% CI 0.9403-0.9991) and extradural hemorrhage (0.9731, 95% CI 0.9113-1.0000) while the worst AUC was on mass effect (0.9216, 95% CI 0.8883-0.9548).

Sensitivities of the algorithms at the high sensitivity operating point for intracranial hemorrhage, calvarial fracture and midline shift were 0.8927 (95% CI 0.8420-0.9315), 0.8718 (95% CI 0.7257-0.9570) and 0.8769 (95% CI 0.7718-0.9453) respectively. Specificities for the same were 0.8427

Finding	**#Positives**	**Sensitivity** (95% **CI)**	**Specificity** (95% **CI)**
Intracranial Hemorrhage	207	0.9807 (0.9513-0.9947)	0.9873 (0.9804-0.9922)
Intraparenchymal	157	0.9809 (0.9452-0.9960)	0.9883 (0.9818-0.9929)
Intraventricular	44	1.0000 (0.9196-1.0000)	1.0000 (0.9979-1.0000)
Subdural	44	0.9318 (0.8134-0.9857)	0.9965 (0.9925-0.9987)
Extradural	27	1.0000 (0.8723-1.0000)	0.9983 (0.9950-0.9996)
Subarachnoid	51	1.0000 (0.9302-1.0000)	0.9971 (0.9933-0.9991)
Fracture	143	1.0000 (0.9745-1.0000)	1.0000 (0.9977-1.0000)
Calvarial Fracture	89	0.9888 (0.9390-0.9997)	0.9947 (0.9899-0.9976)
Midline Shift	54	0.9815 (0.9011-0.9995)	1.0000 (0.9979-1.0000)
Mass Effect	132	0.9773 (0.9350-0.9953)	0.9933 (0.9881-0.9967)

(a) Qure25k dataset: performance of the NLP algorithm in inferring findings from the reports. This is measured on 1779 reports from the Qure25k dataset

Finding	**Reader 1 & 2**		**Reader 2 & 3**		**Reader 3 & 1**		**All**
	Agreement %	Cohen's κ	Agreement %	Cohen's κ	Agreement %	Cohen's κ	Fleiss' κ
Intracranial hemorrhage	89.00	0.7772	90.84	0.8084	88.39	0.7646	0.7827
Intraparenchymal	91.24	0.7865	90.63	0.7651	90.84	0.7719	0.7746
Intraventricular	96.13	0.7042	97.15	0.7350	95.72	0.6550	0.6962
Subdural	87.98	0.4853	93.08	0.6001	90.02	0.5624	0.5418
Extradural	97.35	0.5058	98.37	0.7251	98.17	0.5995	0.6145
Subarachnoid	93.08	0.6778	90.84	0.6058	90.84	0.6363	0.6382
Calvarial Fracture	91.85	0.5771	92.06	0.3704	91.24	0.3637	0.4507
Midline shift	88.19	0.5804	87.17	0.5344	93.69	0.7036	0.5954
Mass effect	86.35	0.6541	87.98	0.6747	86.97	0.6837	0.6698

(b) CQ500 dataset: concordance between the readers

Table 3: Reliability of the gold standards for Qure25k and CQ500 datasets. On Qure25k, we used an NLP algorithm to infer findings from a radiologist's report. Three radiologists reviewed each of the 491 cases on CQ500 dataset and majority vote of the readers is used as gold standard. Table 3a shows the estimates of reliability of the used NLP algorithm while Table 3b shows the reliability and concordance of radiologists' reads

Finding	**AUC** (95% **CI)**	**High sensitivity operating point**		**High specificity operating point**	
		Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Intracranial hemorrhage	0.9194 (0.9119-0.9269)	0.8990 (0.8865-0.9105)	0.7362 (0.7298-0.7425)	0.8309 (0.8156-0.8454)	0.9059 (0.9016-0.9100)
Intraparenchymal	0.8977 (0.8884-0.9069)	0.8932 (0.8789-0.9064)	0.6501 (0.6433-0.6569)	0.7670 (0.7479-0.7853)	0.9046 (0.9003-0.9087)
Intraventricular	0.9559 (0.9424-0.9694)	0.8922 (0.8592-0.9197)	0.9601 (0.9574-0.9627)	0.9220 (0.8927-0.9454)	0.9267 (0.9231-0.9302)
Subdural	0.9161 (0.9001-0.9321)	0.8881 (0.8588-0.9131)	0.7246 (0.7184-0.7307)	0.7816 (0.7448-0.8153)	0.9135 (0.9096-0.9173)
Extradural	0.9288 (0.9083-0.9494)	0.8690 (0.8246-0.9056)	0.8210 (0.8157-0.8262)	0.8207 (0.7716-0.8631)	0.9068 (0.9027-0.9107)
Subarachnoid	0.9044 (0.8882-0.9205)	0.8756 (0.8468-0.9007)	0.7430 (0.7369-0.7489)	0.7643 (0.7286-0.7974)	0.9079 (0.9039-0.9119)
Calvarial fracture	0.9244 (0.9130-0.9359)	0.8982 (0.8777-0.9163)	0.7797 (0.7739-0.7854)	0.8115 (0.7857-0.8354)	0.9020 (0.8978-0.9061)
Midline Shift	0.9276 (0.9139-0.9413)	0.8979 (0.8724-0.9198)	0.8698 (0.8652-0.8744)	0.8664 (0.8381-0.8913)	0.9056 (0.9015-0.9096)
Mass Effect	0.8583 (0.8462-0.8703)	0.8622 (0.8439-0.8792)	0.6157 (0.6089-0.6226)	0.7086 (0.6851-0.7314)	0.9068 (0.9026-0.9108)

(a) Qure25k dataset: the algorithms' performance

Finding	**AUC** (95% **CI)**	**High sensitivity operating point**		**High specificity operating point**	
		Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Intracranial hemorrhage	0.9419 (0.9187-0.9651)	0.8927 (0.8420-0.9315)	0.8427 (0.7952-0.8829)	0.8049 (0.7439-0.8568)	0.9161 (0.8777-0.9455)
Intraparenchymal	0.9544 (0.9293-0.9795)	0.8955 (0.8309-0.9417)	0.8655 (0.8257-0.8992)	0.8433 (0.7705-0.9003)	0.9076 (0.8726-0.9355)
Intraventricular	0.9310 (0.8654-0.9965)	0.8929 (0.7177-0.9773)	0.9179 (0.8891-0.9413)	0.8571 (0.6733-0.9597)	0.9266 (0.8989-0.9486)
Subdural	0.9521 (0.9117-0.9925)	0.8868 (0.7697-0.9573)	0.8813 (0.8472-0.9100)	0.8868 (0.7697-0.9573)	0.9041 (0.8726-0.9300)
Extradural	0.9731 (0.9113-1.0000)	0.8462 (0.5455-0.9808)	0.8933 (0.8621-0.9195)	0.8462 (0.5455-0.9808)	0.9121 (0.8831-0.9359)
Subarachnoid	0.9574 (0.9214-0.9934)	0.8833 (0.7743-0.9518)	0.8794 (0.8448-0.9086)	0.8667 (0.7541-0.9406)	0.9049 (0.8732-0.9309)
Calvarial fracture	0.9624 (0.9204-1.0000)	0.8718 (0.7257-0.9570)	0.9027 (0.8715-0.9284)	0.8718 (0.7257-0.9570)	0.9027 (0.8715-0.9284)
Midline shift	0.9697 (0.9403-0.9991)	0.8769 (0.7718-0.9453)	0.9131 (0.8823-0.9381)	0.9231 (0.8295-0.9746)	0.9038 (0.8717-0.9300)
Mass effect	0.9216 (0.8883-0.9548)	0.8898 (0.8220-0.9384)	0.7830 (0.7370-0.8242)	0.8110 (0.7320-0.8750)	0.9093 (0.8750-0.9368)

(b) CQ500 dataset: the algorithms' performance

Table 4: Performance of the algorithms on the Qure25k and CQ500 datasets. Neither of the datasets was used during the training process. AUCs are shown for 9 critical CT findings on both these datasets. Two operating points were chosen on the ROC for high sensitivity and high specificity respectively.

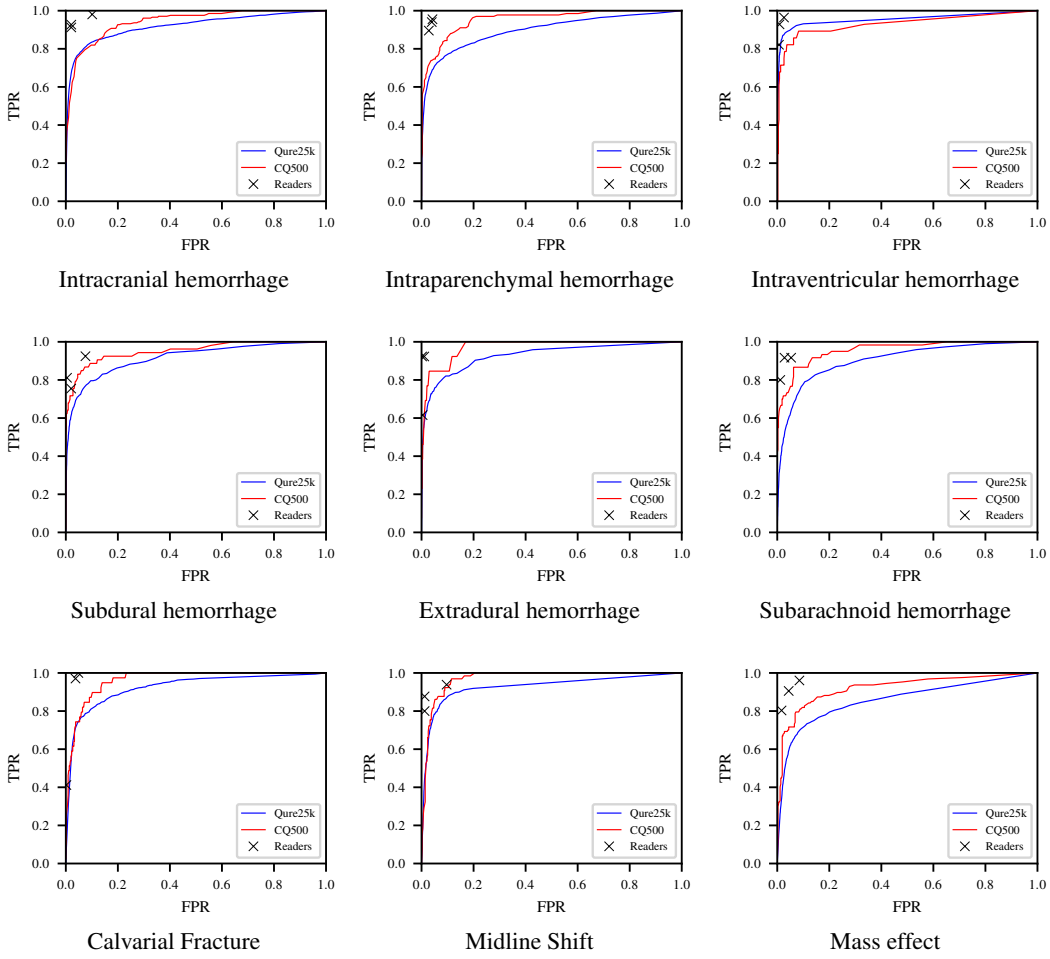


Figure 3: Receiver operating characteristic (ROC) curves for the algorithms on Qure25k and CQ500 datasets. Blue lines are for the Qure25k dataset and Red lines are for the CQ500 dataset. Readers' TPR and FPR against consensus on CQ500 dataset are plotted along with the ROCs for comparison

(95% CI 0.7952-0.8829), 0.9027 (95% CI 0.8715-0.9284) and 0.9131 (95% CI 0.8823-0.9381) respectively.

Algorithms' performance difference between the CQ500 dataset and the Qure25k dataset was most amplified for mass effect (AUC of 0.9216 vs 0.8583) and the least for intracranial hemorrhage (AUC of 0.9419 vs 0.9149).

4 Discussion

Automated and semi automated detection of findings from head CT scans have been studied by other groups. Grewal et al. [11] developed a deep learning approach to automatically detect intracranial hemorrhages. They reported a sensitivity of 0.8864 and a positive predictive value (precision) of 0.8125 on a dataset of 77 brain CT scans read by three radiologists. However, the types of intracranial hemorrhage considered were not mentioned in their report. Traditional computer vision techniques like morphological processing were used by Zaki et al. [53] and Yamada et al. [54] to detect and retrieve scans with fracture respectively. Neither of the two studies measured their accuracies on a clinical dataset. Automated midline shift detection was also explored[55, 56, 57] using non deep learning methods. Convolutional neural networks were used by Gao et al. [10] to classify head CT scans to help diagnose Alzheimer's disease. More recently, Prevedello et al. [58] evaluated the

Gold standard / **Finding**	Reader 1	Reader 2	Reader 3	Majority Vote
Intracranial hemorrhage	0.9080	0.9413	0.9356	0.9419
Calvarial fracture	0.9198	0.8653	0.9523	0.9624
Midline shift	0.9545	0.9386	0.9461	0.9697

Table 5: AUCs of the algorithms on CQ500 dataset when a single reader is considered the gold standard and when the majority vote is considered the gold standard

performance of a deep learning algorithm on a dataset of 50 scans to detect hemorrhage, mass effect, or hydrocephalus (HMH) and suspected acute infarct (SAI). They reported AUCs of 0.91 and 0.81 on HMH and SAI respectively.

Our study is the first to describe the development of a system that separately identifies critical abnormalities on head CT scans and to conduct a validation with a large number (21095) of scans sampled uniformly from the population distribution. We also report the algorithms' accuracy versus a consensus of 3 radiologists on a second independent dataset, the CQ500 dataset. We make this dataset and the corresponding reads available for public access, so that they can be used to benchmark comparable algorithms in the future. Such publicly available datasets had earlier spurred comparison of the algorithms in other tasks like lung nodule detection[59] and chest radiograph diagnosis[8].

This work is novel because it is the first large study describing the use of deep learning on head CT scans to detect and separately report accuracy on each finding, including the five types of intracranial hemorrhage. Further, there is very little literature to date describing the accurate use of deep learning algorithms to detect cranial fractures - we demonstrate that deep learning algorithms are able to perform this task with high accuracy. The clinical validation of algorithms that detect mass effect and midline shift (both used to estimate severity of a variety on intracranial conditions and the need for urgent intervention) on such a large number of patients is also unique.

The algorithms produced fairly good results for all the target findings on both the Qure25k and CQ500 datasets. AUCs for all the findings except mass effect were greater than or approximately equal to 0.9. AUCs on CQ500 dataset were significantly better than that on the Qure25k dataset. We hypothesize that this might be because of two reasons. Firstly, since radiologists reading Qure25k dataset had access to clinical history of the patient, their reads incorporated extra clinical information not available in the scans. The algorithms didn't have access to this information and therefore did not perform well. Secondly, majority vote of three readers is a better gold standard than that of one reader. Indeed, we observed that AUCs of the algorithms on CQ500 were lower when a single reader was considered the gold standard instead of the majority vote (see Table 5).

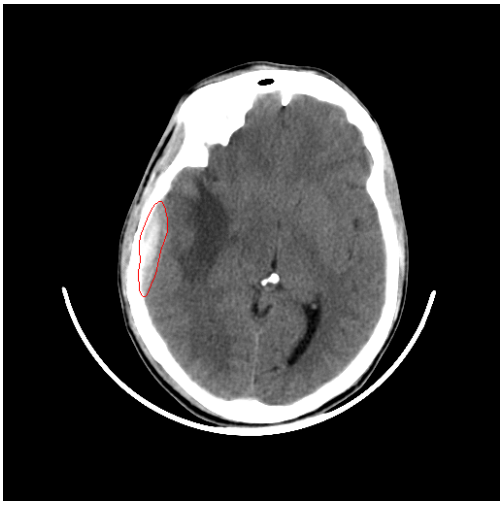
We have done an informal qualitative analysis by going through the algorithms' output and the three readers' opinions on the CQ500 dataset. It was observed that the algorithms produced good results for normal scans without bleed, scans with medium to large sized intraparenchymal and extraaxial hemorrhages, hemorrhages with fractures and in predicting midline shift. There was room for improvement for small sized intraparenchymal, intraventricular hemorrhages and hemorrhages close to skull base. In this study, we didn't separate chronic and acute hemorrhages. This resulted in occasional prediction of scans with infarcts and prominent CSF spaces as intracranial hemorrhages. Future research can improve the algorithms to mitigate these failure modes. We show some accurate and erroneous predictions of the algorithms in Figure 5.

Our study has several limitations. Although the selection strategy ensured that there were a significant number of positive scans in the CQ500 dataset for most of our target findings, number of scans with extradural hemorrhage were found only to be 13. This makes the results of extradural hemorrhage on the CQ500 dataset somewhat statistically insignificant.

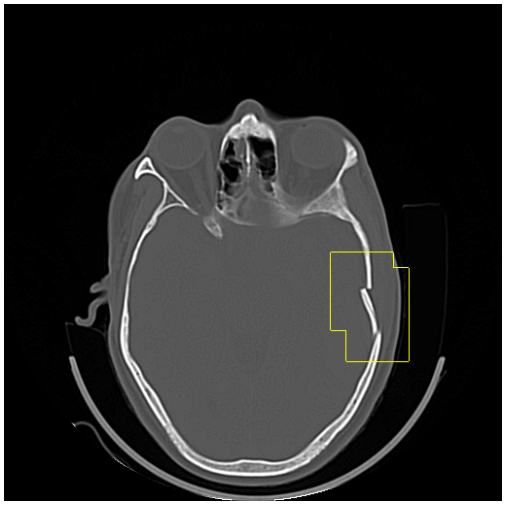
For the scans in the CQ500 dataset, concordance between the three radiologists was not very high for all the findings (see Table 3b). In particular, fracture had low Cohen's kappas of 0.5771, 0.3704 and 0.3637 between the pair of readers. This might be because of non-availability of clinical history to the readers. We observed that the readers were either very sensitive or very specific to a particular finding. For example, two readers were highly sensitive to calvarial fracture while the third reader

Finding	Reader 1		Reader 2		Reader 3	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Intracranial hemorrhage	0.9805	0.8986	0.9268	0.9790	0.9122	0.9790
Calvarial fracture	1.0000	0.9519	0.9706	0.9628	0.4118	0.9978
Midline shift	0.8769	0.9883	0.9385	0.9038	0.8000	0.9883

Table 6: Sensitivities and specificities of the readers versus the majority vote



(a) Output produced by the hemorrhage segmentation algorithms.



(b) Localization output produced by the calvarial fracture detection algorithm.

Figure 4: Localizations produced by the algorithms. These can provide a visual display of the results.

was highly specific. Sensitivities and specificities of the readers for intracranial hemorrhage, calvarial fracture and midline shift are shown in Table 6.

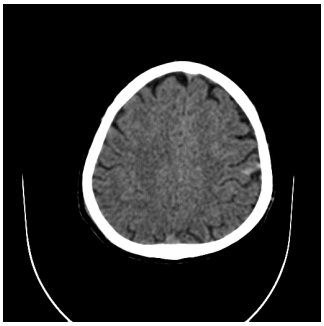
Another limitation of our study is that we have not excluded follow up scans of a patient from the CQ500 dataset. This is primarily due to the fact that there were very few scans reported with some of our target abnormalities like extradural and intraventricular hemorrhages. We couldn't present the extent of this limitation because of non-availability of unique identifier of the patients in this dataset. However, note that there was no overlap of patients or scans between the CQ500 dataset and the development dataset.

Since we also trained segmentation networks for hemorrhage detection algorithms, we can also output a mask representing the precise location and extent of the hemorrhage (except for the subarachnoid hemorrhage), in addition to detecting its presence. Similarly, our calvarial fracture detection algorithm can produce a localization output. These outputs are represented in Figure 4. However in this paper, we have not quantified their accuracy through independent radiologist reads. In this study, we have limited our algorithm to the calvarial (cranial vault) fractures; the success of the algorithm on detecting these means that our ongoing research can include an extension of the algorithm to all other cranial and facial fractures.

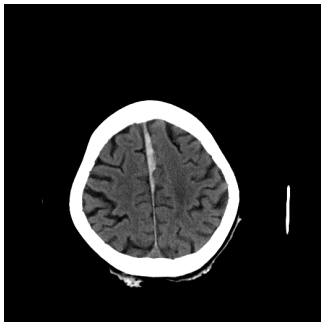
5 Conclusion

Our results show that deep learning algorithms can be trained to detect critical findings from head CT scans with good accuracy. The strong performance of deep learning algorithms suggest they could be a helpful adjunct for identification of acute Head CT finding in a trauma setting, providing a lower performance bound for quality and consistency of radiologic interpretation. It could also be feasible to automate the triage process of Head CT scans with these algorithms. However, further research is necessary to determine if these algorithms enhance the radiologists' efficiency and ultimately improve patient care and outcome.

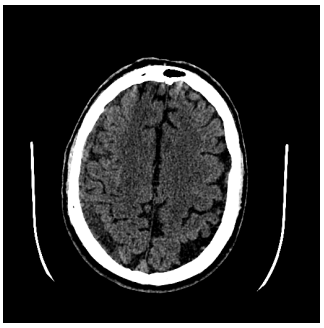
(a) Intraparenchymal hemorrhage (in left frontal region)



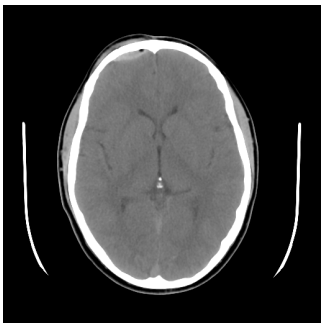
(b) Subarachnoid hemorrhage (in left parietal region)



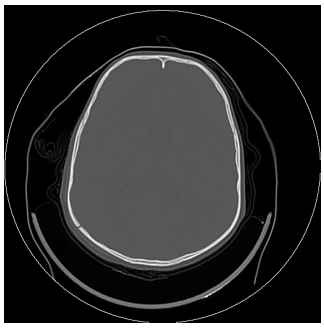
(c) Subdural hemorrhage (along falx)



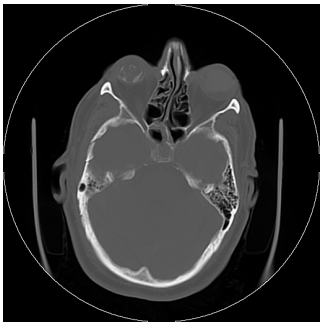
(d) Subdural hemorrhage (chronic in right parietal convexity)



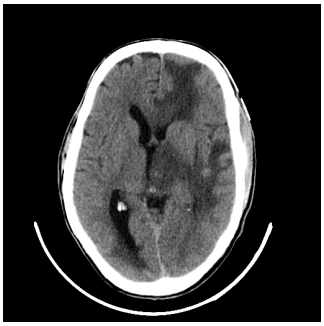
(e) Extradural hemorrhage (in right frontal convexity)



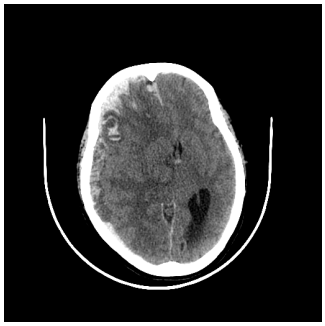
(f) Calvarial fracture (in left parietal bone)



(g) Calvarial fracture (in right temporal bone)



(h) Midline shift

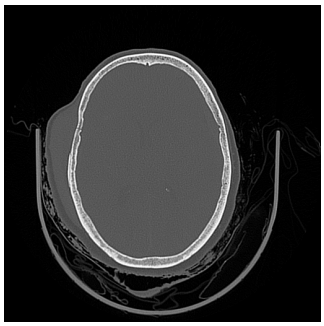


(i) Midline shift

Accurate predictions: True Positives

(j) False negative: tiny intraventricular hemorrhage

(k) False positive: predicted as subdural hemorrhage



(l) False negative: Calvarial fracture (in right parietal bone)

Erroneous predictions

Figure 5: Some accurate and erroneous predictions of the algorithms

References

[1] Helmut Ringl, Ruediger E Schernthaner, Gerd Schueller, Csilla Balassy, Daniela Kienzl, Ana Botosaneanu, Michael Weber, Christian Czerny, Stefan Hajdu, Thomas Mang, et al. The skull unfolded: a cranial ct visualization algorithm for fast and easy detection of skull fractures. *Radiology*, 255(2):553–562, 2010.

[2] J. P. Coles. Imaging after brain injury. *BJA: British Journal of Anaesthesia*, 99(1):49–60, 2007. doi: 10.1093/bja/aem141. URL +http://dx.doi.org/10.1093/bja/aem141.

[3] David B Larson, Lara W Johnson, Beverly M Schnell, Shelia R Salisbury, and Howard P Forman. National trends in ct use in the emergency department: 1995–2007. *Radiology*, 258(1): 164–173, 2011.

[4] Linda Papa, Ian G Stiell, Catherine M Clement, Artur Pawlowicz, Andrew Wolfram, Carolina Braga, Sameer Draviam, and George A Wells. Performance of the canadian ct head rule and the new orleans criteria for predicting any traumatic intracranial injury on computed tomography in a united states level i trauma center. *Academic Emergency Medicine*, 19(1):2–10, 2012.

[5] Michael G Wysoki, Carlos J Nassar, Robert A Koenigsberg, Robert A Novelline, Scott H Faro, and Eric N Faerber. Head trauma: Ct scan interpretation by radiology residents versus staff radiologists. *Radiology*, 208(1):125–128, 1998.

[6] William K Erly, William G Berger, Elizabeth Krupinski, Joachim F Seeger, and John A Guisto. Radiology resident evaluation of head ct scan orders in the emergency department. *American journal of neuroradiology*, 23(1):103–107, 2002.

[7] Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE transactions on medical imaging*, 36 (9):1876–1886, 2017.

[8] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

[9] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[10] Xiaohong W Gao, Rui Hui, and Zengmin Tian. Classification of ct brain images based on deep learning networks. *Computer methods and programs in biomedicine*, 138:49–56, 2017.

[11] Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. *arXiv preprint arXiv:1710.04934*, 2017.

[12] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.

[13] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[15] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1207–1216, 2016.

[16] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, 6:24454, 2016.

[17] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013.

[18] Shu Liao, Yaozong Gao, Aytekin Oto, and Dinggang Shen. Representation learning: a unified deep learning framework for automatic prostate mr segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 254–261. Springer, 2013.

[19] Jay Patravali, Shubham Jain, and Sasank Chilamkurthy. 2d-3d fully convolutional neural networks for cardiac mr segmentation. *arXiv preprint arXiv:1707.09813*, 2017.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

[21] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852. IEEE, 2017.

[22] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[23] Centers for Disease Control, Prevention, et al. Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report*, 52(Suppl. 1):1–17, 2003.

[24] Sebastian Harth, Martin Obert, Frank Ramsthaler, Christina Reuß, Horst Traupe, and Marcel A Verhoff. Estimating age by assessing the ossification degree of cranial sutures with the aid of flat-panel-ct. *Legal Medicine*, 11:S186–S189, 2009.

[25] Sébastien Jodogne, Claire Bernard, Magali Devillers, Eric Lenaerts, and Philippe Coucke. Orthanc-a lightweight, restful dicom server for healthcare and medical research. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 190–193. IEEE, 2013.

[26] Open Health Imaging Foundation. Ohif viewer. URL `http://ohif.org`.

[27] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[28] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[29] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

[30] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[31] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[37] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[38] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[40] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231, 2013.

[41] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[42] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[43] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[44] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[45] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[46] William E Brant and Clyde A Helms. *Fundamentals of diagnostic radiology*. Lippincott Williams & Wilkins, 2012.

[47] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[48] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[49] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[50] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. SciPy society Austin, 2010.

[53] Wan Mimi Diyana Wan Zaki, Mohammad Faizal Ahmad Fauzi, and Rosli Besar. A new approach of skull fracture detection in ct brain images. In *International Visual Informatics Conference*, pages 156–167. Springer, 2009.

[54] Ayumi Yamada, Atsushi Teramoto, Tomoko Otsuka, Kohei Kudo, Hirofumi Anno, and Hiroshi Fujita. Preliminary study on the automated skull fracture detection in ct images using black-hat transform. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 6437–6440. IEEE, 2016.

[55] Wenan Chen, Ashwin Belle, Charles Cockrell, Kevin R Ward, and Kayvan Najarian. Automated midline shift and intracranial pressure estimation based on brain ct images. *Journal of visualized experiments: JoVE*, (74), 2013.

[56] Huan-Chih Wang, Shih-Hao Ho, Furen Xiao, and Jen-Hai Chou. A simple, fast and fully automated approach for midline shift measurement on brain computed tomography. *arXiv preprint arXiv:1703.00797*, 2017.

[57] Furen Xiao, Chun-Chih Liao, Ke-Chun Huang, I-Jen Chiang, and Jau-Min Wong. Automated assessment of midline shift in head injury patients. *Clinical neurology and neurosurgery*, 112 (9):785–790, 2010.

[58] Luciano M Prevedello, Barbaros S Erdal, John L Ryu, Kevin J Little, Mutlu Demirer, Songyue Qian, and Richard D White. Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology*, 285(3):923–931, 2017.

[59] Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.