medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Accuracy of Combined Deep Learning Algorithms in Detecting Spontaneous Intracranial Hemorrhage on Emergent Head CT Scans

Takala Juuso,[1] MD, Peura Heikki,[1] MD, Riku Pirinen, [1] BM, Väätäinen Katri,[2] MD, Sergei Terjajev,[2] MD, Ziyuan Lin,[3] MSc, Rahul Raj,[1] MD, PhD, Korja Miikka,[1] MD, PhD

[1]Department of Neurosurgery, University of Helsinki and Helsinki University Hospital, P.O. Box 266, FI-00029 Helsinki, Finland

[2]Diagnostic Center, Helsinki University Hospital, P.O. Box 266, FI-00029 Helsinki, Finland

[3]Planmeca, Asentajankatu 6, 00880 Helsinki, Finland

Correspondence to Dr. Juuso Takala, Department of Neurosurgery, University of Helsinki, P.O. Box 266, FI-00029 Helsinki, Finland; E-mail address: juuso.takala@helsinki.fi

Word count manuscript: 5921

Short title: Deep learning in intracranial hemorrhage detection

Keywords: Intracranial Hemorrhage; Subarachnoid Hemorrhage; Intracerebral Hemorrhage; Intraventricular Hemorrhage; Computed Tomography; Deep Learning; Convolutional Neural Networks; Computer Vision

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Background: Spontaneous intracranial hemorrhages are life-threatening conditions that require fast and accurate diagnosis. We hypothesized that deep learning (DL) could be utilized to detect these hemorrhages with a high accuracy.

Methods: We developed a DL solution for detecting spontaneous intracerebral (ICH), intraventricular (IVH) and subarachnoid hemorrhages (SAH) from head non-contrast CT (NCCT) scans. The solution included four convolutional neural network (CNN) base models for different hemorrhage types and a CNN metamodel that was trained on top of the base models. We validated the performance of the solution by using a retrospective real-world dataset of consecutive emergency head NCCTs imaged during a 3-month period in 10 different hospitals. The head NCCTs with hemorrhages were stratified into groups by delay from symptom onset to NCCT imaging to better evaluate the suitability of the solution for emergency use.

Results: The real-world validation dataset included 7797 emergency head NCCTs that were imaged between October 1st and December 31st 2021. Of these, 118 were reported to show spontaneous intracranial hemorrhages by on-call radiologists, and 7679 were reported negative for hemorrhages. The developed solution detected all reported 78 (sensitivity 100%) spontaneous intracranial hemorrhages if the head NCCT was presumably or confirmedly taken within 12 hours of symptom onset. When assessed for hemorrhages imaged 12 to 24 hours after symptom onset (13 cases), the sensitivity was 76.5 %. Overall sensitivity for detecting spontaneous intracranial hemorrhages on head NCCTs that were imaged with any delay from symptom onset was 89.8 %, and specificity was 89.5 %. The solution also detected five cases that were missed by on-call radiologists.

Conclusions: The DL solution showed high sensitivity for detecting spontaneous ICHs, IVHs and SAHs within the same time window in which also modern CT scanners work best for detecting acute blood on head NCCTs.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Non-standard Abbreviations and Acronyms

DL = Deep learning

ICH = Intracerebral hemorrhage

IVH = Intraventricular hemorrhage

SAH = Subarachnoid hemorrhage

NCCT = Non-contrast CT

CNN = Convolutional neural network

HUH = Helsinki University Hospital

HU = Hounsfield unit

2D = Two-dimensional

GPU = Graphics processing unit

TTA = Test time augmentation

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Spontaneous intracerebral (ICH) and subarachnoid hemorrhage (SAH) account for approximately one-third of all strokes[1]. Both ICH and SAH are associated with high morbidity and mortality rates [2,3]. Due to the high disease burden of ICH and SAH, prompt and accurate diagnosis as well as quickly initiated therapeutic actions are important[4,5]. For example, spontaneous ICH and SAH have a high risk of rebleed and hematoma expansion, which further worsen the prognosis[6,7]. Non-contrast head computed tomography (NCCT) has a high sensitivity for detecting acute intracranial blood within 12 hours from symptom onset, and therefore the diagnosis of acute intracranial hemorrhages is based on emergent NCCTs [8,9].

The number of medical imaging studies is increasing globally [10–12]. At the same time, there are increasing concerns regarding fatigue of radiologists and its effect on diagnostic accuracy[13,14]. Therefore, new technological solutions assisting clinicians and radiologists in interpreting imaging studies rapidly and accurately could alleviate this issue. Based on these premises, we aimed to develop a novel solution of multiple deep learning (DL) algorithms that would detect spontaneous intracranial hemorrhages, namely ICH, SAH and intraventricular hemorrhage (IVH), with a high accuracy. Given the inherent diagnostic limitation of modern head CT scanners in identifying subacute blood (accuracy highest in early imaging) [8,15], we tried to train the solution to give optimal results when applied to cases imaged within 12 hours from symptom onset.

Methods

Ethical considerations

The local institutional review board of Helsinki University Hospital (HUH) approved the retrospective data collection and study design and granted a waiver for acquiring informed consents (HUS/365/2017; HUS/163/2019; HUS/190/2021). According to the Finnish

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

legislation, no ethics committee approval is needed for retrospective studies that utilize registry or archive data. We gathered all imaging data for algorithm training from the HUH, which consists of 23 separate hospitals. All five Finnish university hospitals, including the HUH, are publicly funded non-profit organizations that provide tertiary health care services for all people living in Finland, regardless of socioeconomic status, insurance status, or race/ethnicity. Therefore, we believe that the collected HUH imaging data for algorithm training and validation is not inherently biased or deliberately discriminative. We conducted the study in line with the Declaration of Helsinki. The proofreading of the text was conducted using Chat GPT 4.0. STARD checklist for this study can be found as an online supplement.

Data availability

Finnish healthcare data for secondary use can be obtained through FINDATA (Social and Health Data Permit Authority according to the Secondary Data Act). The used healthcare data cannot be shared openly. Our solution included four base U-Nets for detecting types of spontaneous intracranial hemorrhages and a metamodel, which was trained on the top of the base U-Nets. We trained independent U-Nets to detect ICH, SAH and IVH (one ICH U-Net, one IVH U-Net and two SAH U-Nets). The training and performance metrics of one of the two SAH U-Nets have been described before [16].

Training data

Training dataset of the three new U-Nets consisted of 63, 50 and 67 head NCCT MPR-reformates (with 512 x 512 dimensions) for ICH, IVH, and SAH, respectively. All patients were imaged and treated at HUH. Segmentations were done using Philips IntelliSpace Discovery [17] and 3D Slicer [18]. Image data and segmentation files were saved and used in

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

NIfTI file format. We used Hounsfield unit (HU) threshold-based method to decrease human error and to increase reproducibility. We determined an acute blood threshold value for each bleeding type; 60-90 HUs for ICH and SAH and 50-90 HUs for IVH. We only segmented the bleeding type of interest for the base model training. To make a final agreement of the segmentation mask, all the segmented scans were reviewed by a cerebrovascular neurosurgeon and/or radiologist. After the review, segmentation masks were saved in a binary format, 0 meaning “no blood” and 1 meaning “blood”. We randomly chose 55 NCCT scans from the base model training NCCTs for the meta-model training dataset. For these 55 NCCTs we segmented all three bleeding types using the same HU thresholds mentioned above. The review process after segmentations remained the same. Eventually, the segmentation masks were saved also in a binary format and each segmentation in its own channel. The final segmentation files for the metamodel training included three channels (one for each bleeding type).

Model architectures and training

The trained three U-Nets were two-dimensional (2D) and had five convolution blocks with residual connections both in the encoder and decoder parts. Dropout and batch normalization layers were also used. Down sampling was done using a 2D convolutional layer with a kernel size of (2,2) and stride of (2,2). Dice loss was used as training loss for ICH and SAH U-Nets. For the IVH U-Net, we used a custom loss function where focal loss and dice loss were combined. Image data for ICH and SAH U-Nets were first clipped between -500 and 500 HUs and then normalized between 0 and 1. Image data for IVH U-Net was clipped between 0 and 200 HUs and then normalized between 0 and 1. Image data and segmentation masks were rotated 90, 180 and 270 degrees as an augmentation method during the training. Before training, each of the datasets were randomly split into training and validation set using 80:20

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

split ratio with 20% of the data reserved for validation purposes during training process. Training was done in Microsoft Azure using NVIDIA Tesla V100 graphics processing unit (GPU) with a batch size of 64, kernel size of (3,3) and dropout rate of 0.2. We used an Adam optimizer with a learning rate of 0,001. The training was set to last 200 epochs and the U-Net with the lowest validation loss value was selected for the final use.

We trained the metamodel after training the base U-Nets. The metamodel included a batch normalization layer and two 2D convolutional layers. The input of the metamodel is a 2D image with 5 channels (original CT slice and predictions from the four base U-Nets). The image channel was clipped between –500 and 500 HUs and normalized between 0 and 1. During the training, image data and corresponding segmentation masks were rotated 90, 180 and 270 degrees as an augmentation method. Similar 80:20 split ratio was used for the metamodel training dataset. Training was done in Microsoft Azure using V100 GPU with batch size of 64, kernel size of 3. The training was set to last 100 epochs. Learning rate of 0.003 and Adam optimizer were used. We used Dice loss as a loss function, and the model with the lowest validation loss was selected for the final use. The model outputs semantic segmentation, i.e. pixel-wise probability of the presence of a hemorrhage in a single NCCT slice.

For the inference pipeline, the combined U-Net solution included a reshaping of image data if the pixel data differed from 512 x 512. After resizing the image, the data was first sent to the base U-Nets with the corresponding normalizations. The base U-Net predictions were saved for later post processing steps. The predictions from the base U-Nets and the original imaging data (with clipping between -500 and 500 HU and normalization) were then directed to the metamodel for the final prediction. We used TensorFlow 2.2.0 to build and train the U-Nets and the metamodel.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Adaptive post-processing steps

To reduce the number of falsely predicted positive findings, we implemented post processing steps. First, we included only predictions in which segmented pixels (i.e. predicted blood) formed a cluster of a minimum 10 pixels in a single slice. A voting support from the base U-Nets was used to direct the slice further for a test time augmentation (TTA) step, or to push the slice forward without TTA. In the voting phase, all predictions from the base U-Nets were summed and compared against the metamodel predictions. If there was an overlap between the metamodel prediction and majority of the base U-Net predictions, the slice was classified as a strong positive. TTA was applied as a post processing step only if the slice was positive, but the prediction was not classified as a strong positive, as described above. By default, the solution made the predictions for input image data without any augmentations. In the TTA step, the image data was flipped both horizontally and vertically. The augmented images were then analyzed, and the predictions were then spatially reversed to match the predictions of the original data. All predictions were then summed and divided by 4. A removal of pixel clusters smaller than 10 pixels was also done after the TTA step. In the final step of the post processing, the positive prediction clusters were combined with the predictions of the base U-Nets, and if the combined cluster size exceeded 125 pixels, the segmentation was classified as a positive prediction. Of note, one 512 x 512 head NCCT slice contains 262 144 pixels. Examples of DL solution's output is presented in figure 1.

Real-world validation dataset

To simulate a real-world emergency imaging setting, we collected a retrospective dataset from 10 different hospitals in the HUH catchment area. These hospitals offer various levels of emergency care, ranging from primary care without neurosurgical services to tertiary care with such services. Together these 10 hospitals cover a catchment area of over 1,700,000

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

inhabitants in Southern Finland. The validation dataset consisted of all consecutive emergency head NCCTs imaged between October 1st and December 31st 2021 and corresponding on-call radiology reports. In more detail, if NCCT scan of adult patients (18 years or older) was performed either with emergent or immediate priority, it was included in the validation dataset. We also collected patient reports of the corresponding emergency room visits and ambulance reports, when available. In addition to these, we gathered information about patient demographics (age, sex, date of birth).

The on-call radiology reports were considered as ground truths, and every report was scrutinized to identify whether on-call radiologists had reported a primary spontaneous intracranial hemorrhage in the scan or not. Primary spontaneous intracranial hemorrhages included non-traumatic aneurysmal and non-aneurysmal SAHs, non-traumatic deep and lobar ICHs, and non-traumatic IVHs. Secondary spontaneous intracranial hemorrhages included SAHs, ICHs or IVHs related to ischemic strokes and tumors, and were excluded from the analysis. Also scans with reported poor quality, wrong imaging protocols (for example contrast-enhanced imaging and angiographies), and wrong reformatting (for example no axial series) were excluded from the dataset. We also excluded NCCT scans with inconclusive reporting of the presence of intracranial hemorrhages, and all follow-up scans. The dataset selection process is presented in figure 2. The dataset did not include head NCCTs from patients who were already admitted to a hospital ward before the NCCT was imaged. Also, if a single patient was imaged multiple times during different emergency clinic visits in the 3 months period, all first scans (excluding follow up scans) were included in the dataset.
The time of the symptom onset and the etiology of hemorrhages were evaluated on the basis of the patient and ambulance reports. If the etiology (traumatic vs. non-traumatic) of the hemorrhage could not be determined, it was classified uncertain. NCCTs with primary spontaneous intracranial hemorrhages were divided into five groups based on the time delay

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

from symptom onset to imaging: 1) the exact time of symptom onset was mentioned in the patient or ambulance report, or in the imaging study referral, and the imaging was carried out within 12 hours from symptom onset, 2) the exact time of symptom onset was not clearly mentioned in the reports, but the imaging was done most likely within 12 hours from symptom onset, and 3) the exact time of symptom onset was not mentioned in the reports, but the imaging was performed most likely within 12- 24 hours from symptom onset or the exact time of symptom onset was mentioned, and imaging was carried out within 12 to 24 hours from symptom onset, 4) the symptom onset was between 24 hours to 7 days before the imaging or 5) the symptom onset was over 7 days before the imaging or the symptom onset remained unclear based on the available reports.

The head NCCT scans in which the on-call radiologist had reported hemorrhage were independently annotated on a slice-level by two study authors (JT and KV) for the presence of hemorrhage. Slices which included 5% or more missing image area (technical shortcoming related to head CT scanners) due to image reformatting were automatically considered negative for hemorrhages. The senior study authors solved any conflicts between the annotators either by removing annotations from the slice or accepting the slice annotations. All imaging studies in the test dataset were in the DICOM format. Nvidia Tesla V100 GPU was also used for running the inference task in Microsoft Azure Machine learning studio.

Statistical analyses

We calculated the patient- and slice-level metrics by using the ground truth radiology reports and Python scripts tailored for these tasks. These reported metrics included sensitivity, specificity, false positive rate, negative predictive value and accuracy. We performed all statistical analyses with the Python package NumPy. We plotted the article figures using Matplotlib Python package.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Technical performance

Figure 3 depicts the time range taken for analyzing all slices in one head NCCT scan (median number of slices 53). The median time of analyzing all slices in a single axial 3 mm NCCT scan was 6.7 seconds (range from 4.2 to 33.1 seconds). The analyzing procedure consisted of multiple steps, such as reading the DICOM files, processing the imaging data, predicting the presence of a hemorrhage, post processing step, and saving the predictions in a NIfTI file.

Validation dataset

The final validation dataset included 118 NCCTs in which on-call radiologists reported spontaneous intracranial hemorrhage and 7679 NCCTs which were reported negative for hemorrhage by the on-call radiologist. The dataset included 4069 NCCTs of women and 3707 NCCTs of men (supplementary material – table 1). The median age of imaged patients was 71 years (range from 18 to 102 years) (supplementary material – table 1). The 7797 NCCTs were imaged using 12 different CT scanners from four vendors (supplementary material – table 2).

Of these, 59 were imaged within 12 hours, 19 presumably within 12 hours and 17 within 24 hours from symptom onset (Table 1). Fourteen intracranial hemorrhages were imaged between 24 hours and 7 days (Table 1). Nine intracranial hemorrhages had a delay more than 7 days, or the symptom onset was unclear (Table 1). Of the 59 early-imaged (within 12 hours) spontaneous intracranial hemorrhages, 32 contained only one type of hemorrhage, 22 two types and 6 three types. Overall, of 118 images in which on-call radiologists reported hemorrhage, 49 included only one type of hemorrhage, 38 included two hemorrhage types and 10 images included three types (supplementary material – table 3). The most common

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

type was ICH (supplementary material – table 3). The dataset also included one acute subdural hemorrhage without known head trauma.

Case-level performance

The developed solution detected hemorrhages in 59 out of 59 patients (sensitivity 100.0 %) who were imaged within 12 hours from symptom onset (Table 1). The post processing step did not rule out any true positive cases. For 19 patients imaged most likely (not 100% certain) within 12 hours from symptom onset, the solution's sensitivity was also 100 % (Table 1). Additional 18 patients were imaged 12 to 24 hours from the symptom onset. The sensitivity of identifying acute hemorrhages among this group was 77.8% [95% confidence interval (CI), 56.3 - 96.6 %] (Table I). The four missed cases included two small ICHs (one in right internal capsule with maximum diameter of 6 mm and one in right pontine region with maximum diameter of 10 mm) and two local SAHs (supplementary material – figures 1.1-1.4). For patients imaged between 24 hours to 7 days, the sensitivity of the solution was 71.4% [95% CI, 47.8 - 95.1 %]. The missed hemorrhage cases included one ICH (left sided thalamic with maximum diameter of 6 mm) and three local SAHs (supplementary material – figures 1.5-1.8). The sensitivity for hemorrhages that were imaged after 7 days from symptom onset or had unclear symptom onset time was 55.6% [95% CI, 23.1 - 88.0 %]. The missed hemorrhage cases included three ICHs (one resorbing left sided ICH near falx with maximum diameter of 5 mm hyperdense remnant remaining, one left sided frontal ICH with maximum diameter of 19 mm, and one right sided ICH in temporal region with maximum diameter of 29 mm) and one local SAH (supplementary material – figures 1.9-1.12). The sensitivity for all 118 cases of spontaneous primary intracranial hemorrhages was 89.8% [95% CI, 84.4 - 95.3 %].

Cases missed clinically but identified by solution

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The solution identified five spontaneous intracranial hemorrhages that were not reported in the initial on-call reports. Of these five cases, one was ICH, two were SAHs, and one was IVH. In one imaging study, the on-call radiologist initially reported that the head NCCT showed a meningioma. Later the report was revised and stated that the finding had hyperdense regions which suggested that the finding was an intracranial bleeding instead of a meningioma. In other cases, the presence of hemorrhage was afterwards reported in additional reports by senior radiologists, or the diagnosis was done by the emergency room clinicians. All identified hemorrhages are presented in the supplementary material (supplementary material – figures 2.1 - 2.5)

Slice-level performance

The overall slice-level sensitivity was 69.3 % [95% CI, 67.3 - 71.3 %] and slice-level specificity 99.6 % [95% CI, 99.6 - 99.6 %]. The slice-level false positive rate for all images in the dataset was 0.4 % [95% CI, 0.4 - 0.4 %]. In more detail, of the 7679 NCCTs reported negative for hemorrhage by the on-call radiologists, the solution predicted falsely positive 1594 slices out of the 408 426 slices (Table 2). Most of the false positive pixel clusters pointed out blood in normal and highly vascularized anatomical structures, e.g. sagittal sinus, cerebellar tentorium, straight sinus, and falx cerebri (supplementary material – figures 3 and 4).

Discussion

Our study presents a novel deep learning solution that can detect three different spontaneous intracranial hemorrhages with a relatively high sensitivity, low false positive rate, and a low processing time per an individual NCCT scan. The solution alarms about an intracranial

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

hemorrhage falsely in approximately every 10th true negative head NCCTs. Most of these false positive findings were present only in few slices and in same anatomical locations. Therefore, these small false positive pixel clusters could be relatively easily assessed as false positive findings by on-call radiologists or clinicians. Structures containing high vascularity or low pressure (slow flowing) blood, such as the choroid plexuses and intracranial sinuses, makes achieving a zero false positive rate a challenging task if a high sensitivity (not missing true acute hemorrhages) is a priority. Most importantly, solutions that have been intended to assist on-call radiologists and clinicians in clinical diagnostics should be less than 100% accurate, as otherwise these solutions can be easily utilized as stand-alone solutions. This was not the aim of our study. When the accuracy drops due to false positive cases, not false negatives, the solution cannot impair the diagnostic accuracy, if used in a clinical setting.

In the previous study [16], we described a deep learning algorithm for detecting SAH from head NCCTs. Despite the algorithm having a high sensitivity for detecting intracranial blood, there was a relatively high number of false positive findings. These false positives were often other intracranial hemorrhage types and various other intracranial pathologies, such as tumors. In brief, even though the sensitivity in detecting blood in head NCCTs was high, the performance of the algorithm in detecting other hemorrhage types was limited and not validated. Due to these facts, algorithm ‘s clinical usability was considered poor, and therefore this new solution for all spontaneous intracranial hemorrhages was developed. The sensitivity of acute intracranial hemorrhage detection of our solution is essentially similar to the reported performance metrics of commercially available solutions [19–22]. However, reliable comparisons between different solutions are difficult to conduct due to the lack of standardized comparison protocols and datasets. Although emergency head NCCTs have a high sensitivity for acute blood [8,9], misidentification of acute intracranial hemorrhage

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

still occurs in clinical settings. Due to potentially life-threating consequences, the false negative interpretations should be minimized. Our solution successfully detected two cases of acute SAHs that were not reported in on-call reports. The missed cases, which were flagged correctly by our solution, further suggest that DL-based solutions could function as peer-readers in parallel with on-call radiologists, improve the diagnostic accuracy in clinical practice, and even save human lives.

Our study has certain acknowledged limitations. First, the final number of true hemorrhage cases in the validation dataset can be considered low. However, the dataset represents a true consecutive patient cohort imaged during a 3-month long period in 10 different hospitals with a catchment area of around 1.7 million inhabitants. Second, the dataset was collected retrospectively and included hospitals from which the original training data was collected. However, we collected all training data prior to October 2021. Since the consecutive validation dataset was collected between October and December 2021, a direct data leak from training dataset to the validation dataset was not possible. Third, the validation dataset was collected from 10 different hospitals, but the validation was still internal. Therefore, the solution's results cannot be generalized outside the study country. Currently, it is still challenging to perform external validations abroad, as very few hospitals have comprehensive datasets simulating real-world scenarios and patient cohorts. Moreover, very few hospitals have capabilities and required research permissions to conduct such validation studies. This is one of the major shortcomings in the field. Fourth, we did not have a possibility to assess our solution's usefulness in the clinical workflow, as the solution is not an officially approved medical device. Fifth, as our aim was not to achieve 100% accuracy, which would increase the likelihood that human experts become replaced by autonomous stand-alone solutions in diagnostic settings, the solution's performance metrics can perhaps be considered satisfactory. Despite the shortcomings, our study may also have a few strengths. U-Nets are

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

widely used for image segmentation tasks, and due to relatively low computational costs of the 2-dimensional U-Net architecture, the solution could be run in a local set up without requiring costly high-end GPUs or central processing units. In this sense, the solution would be scalable also for hospitals with limited internet connections or without access to modern cloud-computing services. Moreover, our study suggests that meta learning can be utilized in combining multiple models without increasing false positive rate. We are not aware of any similar solutions designed for clinical imaging diagnostics.

Conclusions

The presented novel solution detects acute (imaging delay <12 hours) spontaneous ICHs, IVHs and SAHs on head NCCTs with a high sensitivity. The described metamodel approach can ease the developing of similar combined solutions with multiple convolutional neural networks. If the results could at some point be externally validated, this solution might be helpful for on-call radiologists and clinicians particularly in ruling out potentially fatal intracranial hemorrhages in emergency setting. Even though the solution is not yet a proved medical device, it could already be used for research purposes and retrospective quality assessments.

Acknowledgements

This work is part of the AI Head Analysis project of the CleverHealth Network ecosystem (https://www.cleverhealth.fi/en/home), and we thank the ecosystem partners for supporting the project.

Sources of Funding

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

JT would like to thank Maire Taponen foundation and HUS neurocenter for research grants. The study project was supported by a grant from the State Research Funds (Helsinki University Hospital). The funders had no role in the design and conduct of the study; in collection, management, analysis, and interpretation of the data; or in preparation, review, or approval of the manuscript.

Disclosures

None

Supplemental Material

Supplementary Tables 1-3

Supplementary Figures 1-4

STARD checklist

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Krishnamurthi R V, Ikeda T, Feigin VL, Rita A, Krishnamurthi V, Ikeda K/, Feigin /. Global, Regional and Country-Specific Burden of Ischaemic Stroke, Intracerebral Haemorrhage and Subarachnoid Haemorrhage: A Systematic Analysis of the Global Burden of Disease Study 2017. *Neuroepidemiology*. 2020;54:171–179.
2. J van Asch CJ, A Luitse MJ, E Rinkel GJ, van der Tweel I, Algra A, M Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. *Lancet Neurol*. 2010;167.
3. Roquer J, Cuadrado-Godia E, Guimaraens L, Conesa G, Rodríguez-Campello A, Capellades J, García-Arnillas MP, Fernández-Candil JL, Avellaneda-Gómez C, Giralt-Steinhauer E, et al. Short- and long-term outcome of patients with aneurysmal subarachnoid hemorrhage. *Neurology*. 2020;95:e1819–e1829.
4. Hostettler IC, Seiffge DJ, Werring DJ. Intracerebral hemorrhage: an update on diagnosis and treatment. *Expert Rev Neurother*. 2019;19:679–694.
5. Steiner T, Juvela S, Unterberg A, Jung C, Forsting M, Rinkel G. European Stroke Organization Guidelines for the Management of Intracranial Aneurysms and Subarachnoid Haemorrhage. *Cerebrovascular Diseases*. 2013;35:93–112.
6. Davis SM, Broderick J, Hennerici M, Brun NC, Diringer MN, Mayer SA, Begtrup K, Steiner T. Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage. *Neurology*. 2006;66:1175–1181.
7. Naidech AM, Janjua N, Kreiter KT, Ostapkovich ND, Fitzsimmons BF, Parra A, Commichau C, Connolly ES, Mayer SA. Predictors and Impact of Aneurysm Rebleeding After Subarachnoid Hemorrhage. *Arch Neurol*. 2005;62:410–416.
8. Perry JJ, Stiell IG, Sivilotti MLA, Bullard MJ, Émond M, Symington C, Sutherland J, Worster A, Hohl C, Lee JS, et al. Sensitivity of computed tomography performed within six hours of onset of headache for diagnosis of subarachnoid haemorrhage: prospective cohort study. *BMJ*. 2011;343.
9. Vincent A, Pearson S, Pickering JW, Weaver J, Toney L, Hamill L, Hurrell M, Than M. Sensitivity of modern multislice CT for subarachnoid haemorrhage at incremental timepoints after headache onset: a 10-year analysis. *Emergency Medicine Journal*. 2021;emermed-2020-211068.
10. Huang CC, Effendi FF, Kosik RO, Lee WJ, Wang LJ, Juan CJ, Chan WP. Utilization of CT and MRI scanning in Taiwan, 2000-2017. *Insights Imaging*. 2023;14.
11. Christiansen NM, Brabrand M, Fløjstrup M, Bech M, Lassen AT, Mogensen CB, Bogh SB. Utilisation and time to performance of diagnostic imaging in patients admitted to Danish emergency departments: a nationwide register-based study from 2007 to 2017. *BMJ Open*. 2023;13.
12. Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, Bowles EJA, Duncan JR, Greenlee RT, Kushi LH, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA*. 2019;322:843–856.
13. Hanna TN, Zygmont ME, Peterson R, Theriot D, Shekhani H, Johnson J-O, Krupinski EA. The Effects of Fatigue From Overnight Shifts on Radiology Search Patterns and Diagnostic Performance. *J Am Coll Radiol*. 2018;15:1709–1716.
14. Stec N, Arje D, Moody AR, Krupinski EA, Tyrrell PN. A systematic review of fatigue in radiology: Is it a problem? *American Journal of Roentgenology*. 2018;210:799–806.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

15. Dubosh NM, Bellolio MF, Rabinstein AA, Edlow JA. Sensitivity of Early Brain Computed Tomography to Exclude Aneurysmal Subarachnoid Hemorrhage: A Systematic Review and Meta-Analysis. *Stroke*. 2016;47:750–755.
16. Thanellas A, Peura H, Lavinto M, Ruokola T, Vieli M, Staartjes VE, Winklhofer S, Serra C, Regli L, Korja M. Development and External Validation of a Deep Learning Algorithm to Identify and Localize Subarachnoid Hemorrhage on CT Scans. *Neurology*. 2023;10.1212/WNL.0000000000201710.
17. IntelliSpace Discovery | Philips.;Available from: https://www.philips.fi/healthcare/product/HC881015/intellispace-discovery
18. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30:1323–1341.
19. Ginat DT. Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. *Neuroradiology*. 2020;62:335–340.
20. Matsoukas S, Chennareddy S, Kalagara R, Scaggiante J, Smith CJ, Bazil MJ, Reford E, Liu K, Delman BN, Selim MH, et al. Pilot Deployment of Viz-Intracranial Hemorrhage for Intracranial Hemorrhage Detection: Real-World Performance in a Stroke Code Cohort. *Stroke*. 2022;53:E418–E419.
21. Heit JJ, Coelho H, Lima FO, Granja M, Aghaebrahim A, Hanel R, Kwok K, Haerian H, Cereda CW, Venkatasubramanian C, et al. Automated Cerebral Hemorrhage Detection Using RAPID. *AJNR Am J Neuroradiol*. 2021;42:273–278.
22. Voter AF, Meram E, Garrett JW, Yu JPJ. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. *J Am Coll Radiol*. 2021;18:1143.

Tables

Table 1 - Case-Level Performance

	TP	TN	FP	FN	Sensitivity % [95% CI]	Specificity % [95% CI]	FPR % [95% CI]	NPV % [95% CI]	PPV % [95% CI]
All Scans	106	6873	806	12	89.8 % [84.4 - 95.3 %]	89.5 % [88.8 - 90.2 %]	10.5 % [9.8 - 11.2 %]	99.8 % [99.7 - 99.9 %]	11.6 % [9.5 - 13.7 %]
Bleed confirmed < 12 hours	59			0	100 % [100 - 100 %]	N/A	N/A	N/A	N/A
Bleed presumably < 12 h	19			0	100 % [100 - 100 %]	N/A	N/A	N/A	N/A
Bleed presumably <24 hours or confirmed 12- 24h	13			4	76.5 % [56.3 - 96.6 %]	N/A	N/A	N/A	N/A
Bleed 24 h - 7 days	10			4	71.4 % [47.8 - 95.1 %]	N/A	N/A	N/A	N/A
Bleed > 7 days or onset time unclear	5			4	55.6 % [23.1 - 88.0 %]	N/A	N/A	N/A	N/A
Negative scans		6873	806		N/A	89.5 % [88.8 - 90.2 %]	10.5 % [9.8 - 11.2 %]	99.8 % [99.7 - 99.9 %]	11.6 % [9.5 - 13.7 %]

TP = true positive, TN = true negative, FP = false positive, FN = false negative, FPR = false positive rate, NPV = negative predictive value, PPV = positive predictive value, N/A = not applicable

Table 2 - Slice-Level Performance

	Slices	TP	TN	FP	FN	Sensitivity % [95% CI]	Specificity % [95% CI]	FPR % [95% CI]	NPV % [95% CI]	PPV % [95% CI]	Accuracy % [95% CI]
Bleed confirmed < 12 hours	3068	788	1961	38	281	73.7 % [71.1 - 76.4 %]	98.1 % [97.5 - 98.7 %]	1.9 % [1.3 - 2.5 %]	87.5 % [86.1 - 88.8 %]	95.4 % [94.0 - 96.8 %]	89.6 % [88.5 - 90.7 %]
Bleed presumably < 12 hours	1053	336	571	11	135	71.3 % [67.3 - 75.4 %]	98.1 % [97.0 - 99.2 %]	1.9 % [0.8 - 3.0 %]	80.9 % [78.0 - 83.8 %]	96.8 % [95.0 - 98.7 %]	86.1 % [84.0 - 88.2 %]
Bleed presumably <24 hours or Confirmed 12- 24h	872	134	673	1	64	67.7 % [61.2 - 74.2 %]	99.9 % [99.6 - 100.0 %]	0.1 % [0.0 - 0.4 %]	91.3 % [89.3 - 93.3 %]	99.3 % [97.8 - 100.0 %]	92.5 % [90.8 - 94.3 %]
Bleed 24h - 7 days	739	98	549	10	82	54.4 % [47.2 - 61.7 %]	98.2 % [97.1 - 99.3 %]	1.8 % [0.7 - 2.9 %]	87.0 % [84.4 - 89.6 %]	90.7 % [85.3 - 96.2 %]	87.6 % [85.2 - 89.9 %]
Bleed > 7 days or onset time unclear	470	39	375	0	56	41.1 % [31.2 - 50.9 %]	100.0 % [100.0 - 100.0 %]	0.0 % [0.0 - 0.0 %]	87.0 % [83.8 - 90.2 %]	100.0 % [100.0 - 100.0 %]	88.1 % [85.2 - 91.0 %]
False Positive	43752	0	42158	1594	0	N/A	96.4 % [96.2 - 96.5 %]	3.6 % [3.5 - 3.8 %]	N/A	N/A	96.4 % [96.2 - 96.5 %]
True Negative	366268	0	366268	0	0	N/A	100.0 % [100.0 - 100.0 %]	0.0 % [0.0 - 0.0 %]	N/A	N/A	100.0 % [100.0 - 100.0 %]
All	416222	1395	412555	1654	618	69.3 % [67.3 - 71.3 %]	99.6 % [99.6 - 99.6 %]	0.4 % [0.4 - 0.4 %]	99.9 % [99.8 - 99.9 %]	45.8 % [44.0 - 47.5 %]	99.5 % [99.4 - 99.5 %]

TP = true positive, TN = true negative, FP = false positive, FN = false negative, FPR = false positive rate, NPV = negative predictive value, PPV = positive predictive value, N/A = not applicable

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures and Figure Legends

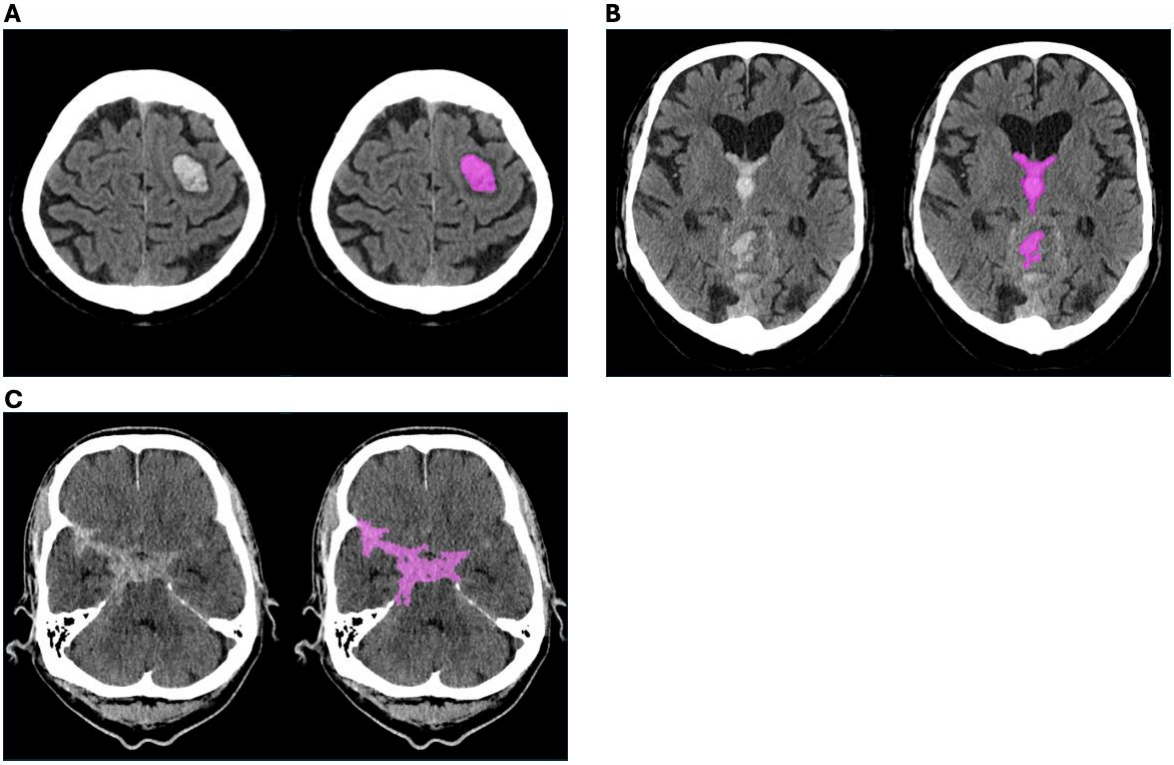

Figure 1 – True Positive Detections of Hemorrhage and Segmentations

True positive detections and corresponding segmentations of A) ICH, B) IVH and C) SAH. The left sided image in each panel shows the original CT scan slice and right sided image displays the DL solution's detection of hemorrhage in the slice (purple color showing detected hemorrhage).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

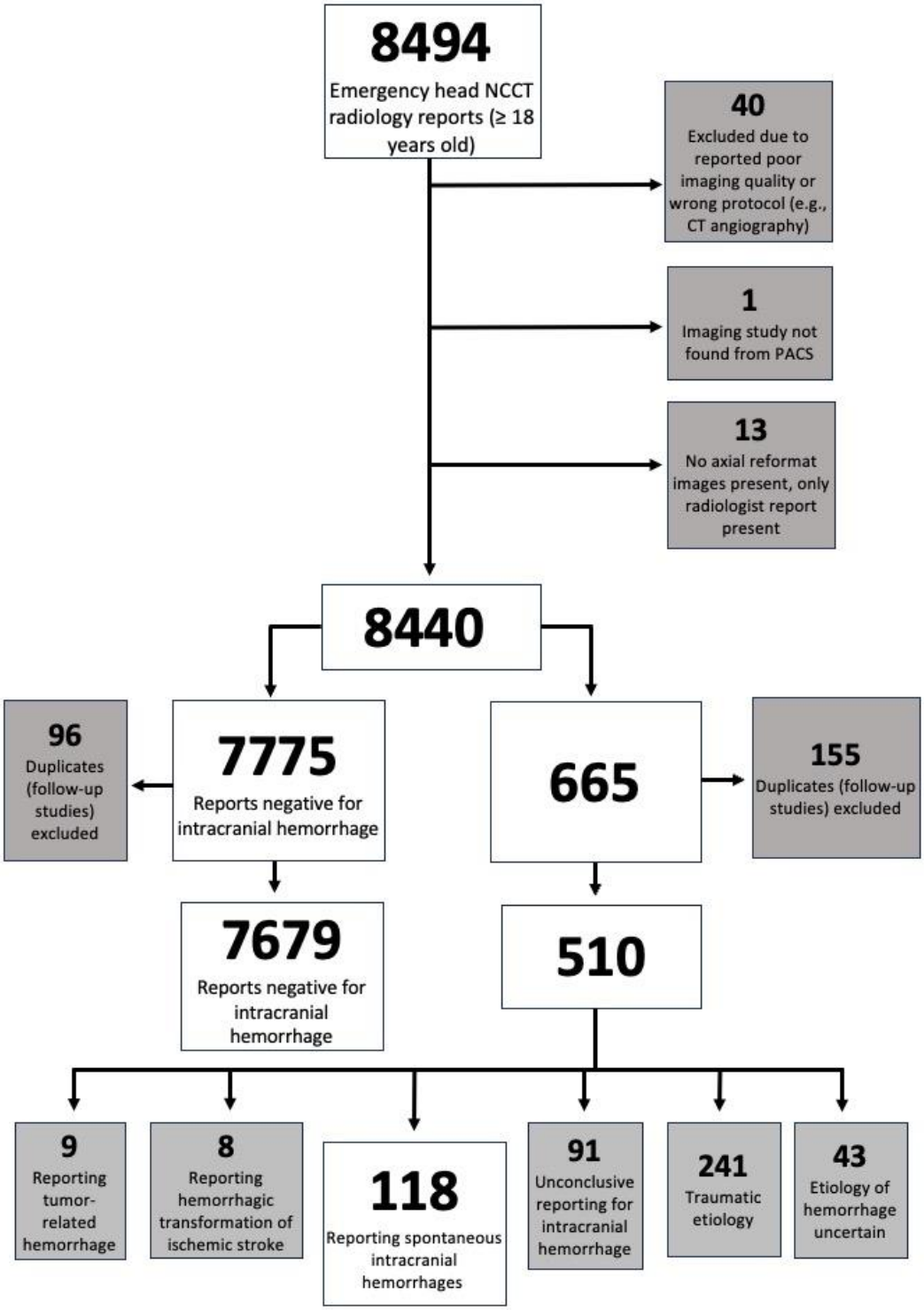

Figure 2 – Dataset Selection Flow Chart

We Reviewed all 8494 on-call report for the NCCTs. The final dataset included 118 NCCTs with spontaneous hemorrhage and 7679 NCCTs which were reported negative for hemorrhage in the on-call report.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

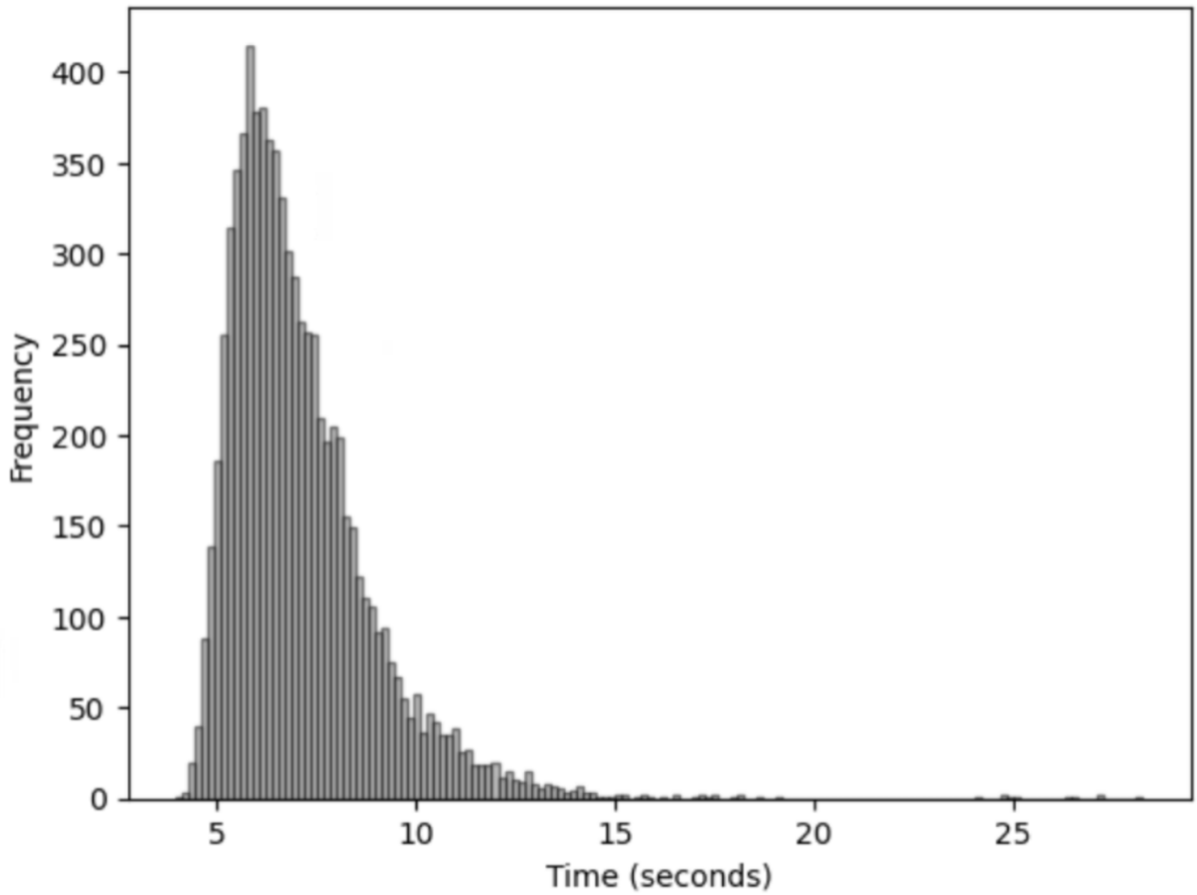

Figure 3 – Deep Learning Solution's Run Times

Histogram of inference times for scans. Median run time for single scan was 6.7 seconds, 95 % of the scans could be analyzed within 10.6 seconds. The run time included processing of all slices in 3 mm axial reformat scan.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.28.24308084; this version posted May 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

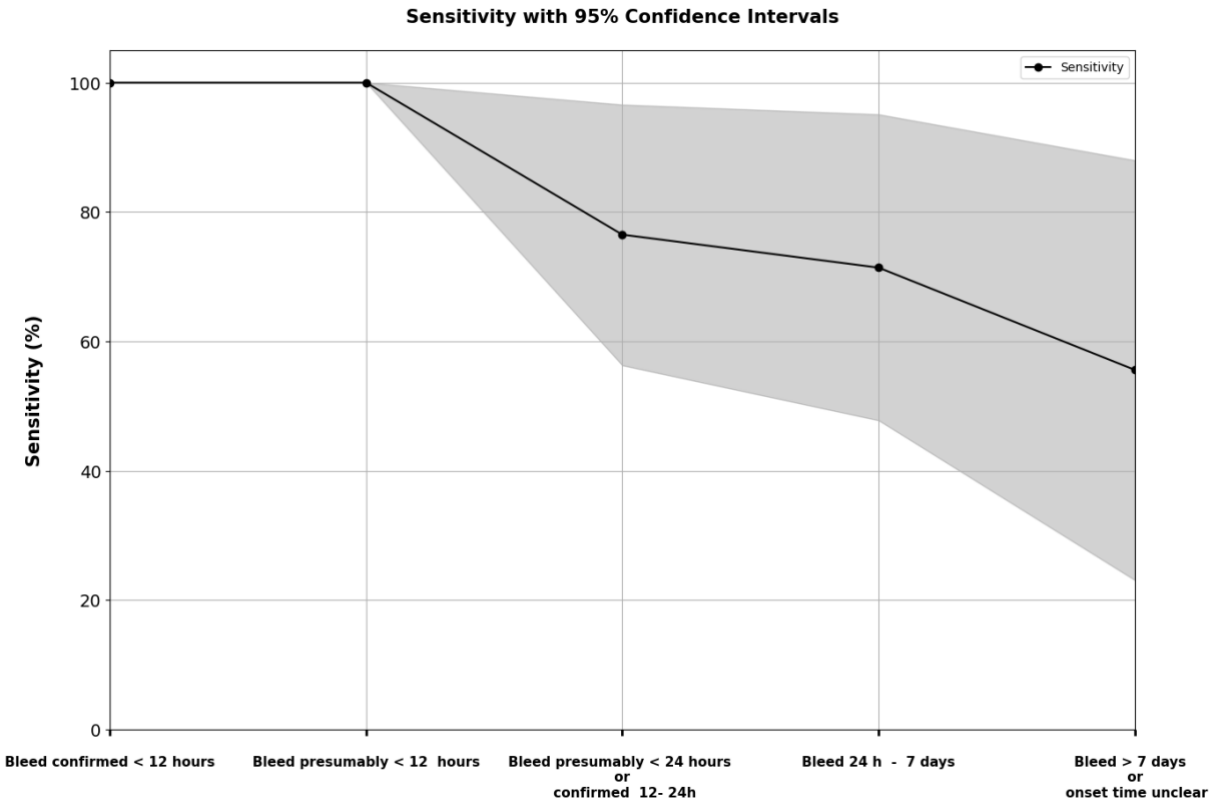

Figure 4 – Deep Learning Solution's Sensitivity

Sensitivity of the deep learning solution for detecting hemorrhages at different time point groups. The shaded are area in the plot represents the 95 % confidence interval.