OPEN ACCESS Freely available online

Journal of Biomedical Engineering and Medical Devices

Research Article

Artefact-tolerant Intracranial Haemorrhage Segmentation Method of Non-contrast CTs: An Open-sourced Tool and Dataset for Algorithm Development

Antonios Konstantinos Thanellas*[1], Mikko Lilja[2], Nik Lygeros[3], Teijo Kottila[4], Miikka Korja[5]

[1]Department of Information Management, Helsinki University Hospital, Paciuksenkatu 25, FI-00270 Helsinki, Finland;[2]Planmed Oy, Helsinki, Finland;[3]LGPC (UMR 5285), Université de Lyon, 69616, Villeurbanne, France;[4]Department of Neuroscience and Biomedical Engineering, Aalto University, Helsinki, Finland;[5]Department of Neurosurgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

ABSTRACT

Objectives: We aimed to create an artefact-tolerant and fully automated segmentation method intended to reduce the workload of medical experts who segment head computed tomography images of intracranial haemorrhage patients.

Methods: We developed a segmentation algorithm that combines 2D and 3D intensity thresholding, morphological operations, and entropy filtering. We tested the algorithm's performance against gold standard segmentations on preoperative and postoperative/posttreatment head computed tomography images of 145 patients with intracranial bleeding. We compared the fully automated algorithm against a simpler thresholded method.

Results: The fully automated algorithm correctly segmented blood in 98.62% of patients, in 2277 out of 2449 positive slices (92.97%), and in 54.12% of positive voxels. It incorrectly segmented blood in 0.63% of patients' negative voxels. The Dice coefficient at voxel level was 0.20.

Conclusion: The open-sourced algorithm may facilitate the segmentation of a wide quality range of preoperative or postoperative/posttreatment head computed tomography scans with intracranial haemorrhage.

Keywords: Segmentation; Head computed tomography; Non-contrast; Subarachnoid haemorrhage; Intracranial haemorrhage

Abbreviations: ICH: Intracerebral Haemorrhage; CT: Head Computed Tomography; AVM: Arteriovenous Malformations; SAH: Subarachnoid Haemorrhage; EDH: Epidural Haemorrhage; SDH: Subdural Haemorrhage; HU: Hounsfield Unit; NCCT: Non Contrast CT; FOV: Field of View; PVE: Partial Volume Effects; PACS: Picture Archiving and Communication System; MPR: Multi Planar Reformat.

INTRODUCTION

Non-traumatic intracerebral haemorrhage, also known as spontaneous Intracerebral Haemorrhage (ICH), is a severe form of stroke accompanied by acute and life-threatening bleeding within the cranium. Brain aneurysms, Arteriovenous Malformations (AVM), and hypertension are among the most important underlying aetiologies of ICH. ICH is associated with high mortality and disability rates and extensive health care costs [1, 2]. The bleeding can break into the subarachnoid space (a condition known as Subarachnoid Haemorrhage or SAH), ventricular space (Intraventricular Haemorrhage or IVH), the space between the inner table of the skull and the dura mater (Epidural Haemorrhage or EDH), and the space between the meningeal layer of the dura and the arachnoid membrane (Subdural Haemorrhage or SDH).

Head Computed Tomography (CT) is the cornerstone of diagnostics, allowing rapid and accurate detection of ICH.

The time needed for interpretation of radiological results is dictated by the medical urgency, the patient's status as inpatient or outpatient, and the available radiology workforce and can vary

Correspondence to: Antonios Konstantinos Thanellas, Department of Information Management, Helsinki University Hospital, Paciuksenkatu 25, FI-00270 Helsinki, Finland, E-mail: antonios.thanellas@hus.fi

Received: 27-Jul-2022, Manuscript No. BEMD-21-002-PreQC-22; **Editor assigned:** 01-Aug-2022, PreQC No. BEMD-21-002-PreQC-22 (PQ); **Reviewed:** 15-Aug-2022, QC No. BEMD-21-002-PreQC-22; **Revised:** 22-Aug-2022, Manuscript No. BEMD-21-002-PreQC-22 (R); **Published:** 29-Aug-2022, DOI: 10.35248/2475-7586.22.08.231

Citation: Thanellas AK, Lilja M, Lygeros N, Kottila T, Korja M (2022) Artefact-tolerant Intracranial Haemorrhage Segmentation Method of Non-contrast CTs: An Open-sourced Tool and Dataset for Algorithm Development. J Biomed Eng & Med Dev. 08:231

markedly [3]. The steady increase in the demand for radiological imaging [4] combined with the shortage of radiologists in the UK [5], Europe and Japan [6] (less in the US [7]) highlights the need for measures to reduce the workload of the current workforce. Software solutions that will assist diagnosis or facilitate and reduce the time needed for a specialist to interpret radiological images have become necessary. In fact, machine learning as well as more classical (non-machine learning) image processing techniques have been employed to address this particular problem. In either case, material that has been segmented by experts is required to train and validate such methods. However, segmentation of the regions of interest by medical professionals has become one of the major hindrances in a large-scale development of clinically useful algorithms. The segmentation task requires locating the exact spot of the lesion at voxel level, and, under the gold standard methodology, experts need to draw every voxel of the area that the lesion occupies, to which an algorithm's performance is then compared [8, 9]. For example, segmenting diffuse SAH is a time-consuming task since the blood can be present in any part of the subarachnoid space and the surface covered can be very extensive compared to other intracranial haemorrhages.

In CT imaging, the Hounsfield Unit (HU) scale is used to express the attenuation of the X-ray beams while passing through the patient to reach the detectors. Assuming that the intensity values of acute blood in a Non-Contrast CT (NCCT) are relatively fixed in the range of 50 ~ 60 HU, with some variations due to different tube voltage and temperature, detecting blood from an NCCT should be straightforward. However, the type of reconstruction filter, the size of the scanned Field of View (FOV), the location of the FOV, Partial Volume Effects (PVE), and noise can lead to an image with observed blood HU values being quite different from the usual ones. Moreover, in a clinical setting NCCTs exhibit high variability, emerging from different causes, which in turn increases the complexity of a correct interpretation. Co-existence of many medical conditions in NCCTs, such as brain tumors [10], contribute to this complexity. Similarly, normal anatomy, such as the brain sinuses, thickened tentorium, calcifications [11-13], and even physiological brain ageing, can create challenges in segmentation and even lead to a misperception of blood. Therefore, better segmentation approaches are needed to reduce the labour-intensive task of creating correctly segmented datasets necessary for the training and testing of machine learning and neural network methods. The purpose of this work is: 1) to present solutions to the challenges faced in the task of the intracranial haemorrhage segmentation when using the highly variable hospital datasets, and 2) to propose a fully automated segmentation algorithm that will assist in the segmentation process of noisy images.

MATERIALS AND METHODS

Patient and control group selection

The search criteria used to collect patient and control data from the Picture Archiving and Communication System (PACS) were encoded with the International Statistical Classification of Diseases and Related Health Problems (ICD-10) medical classification list. Code I60 along with all of its sub-codes I60.xx were used for patient group retrieval. The category code I60 translates to subarachnoid haemorrhage diagnosis, while accompanying digits describe the cause of bleed in more detail.

The control group was retrieved by searching for subjects imaged (head CT scan) due to headache and who had no findings of acute intracranial haemorrhages. The codes R51 and G44.2 were used for control group retrieval. The R51 category translates to headache and G44.2 to a tension-type headache.

Image data

The image data collected from 2011 to 2018 consisted of reconstructed non-source Multi-Planar Reformat (MPR) volumes. The Digital Imaging and Communications in Medicine (DICOM) images were converted to the Neuroimaging Informatics Technology Initiative format (NIFTI-1) using the dcm2niix utility. The patient group consisted of 145 MPR volumes, with the number of unique patients being 142; one patient had two and another patient one follow-up scans. The control group consisted of 150 MPR volumes that were randomly picked from a larger control cohort, with the number of unique patients equaling 150. The clinical findings of the patients are shown in Table 1.

Metadata	Frequency [%]
Clinical Status Preoperative	40 [27.59]
Postoperative	64 [44.14]
Posttreatment (endovascular)	34 [23.45]
Ventriculostomy	7 [4.83]
Bleeding Type Subarachnoid	127 [87.59]
Intraventricular	103 [71.03]
Intracerebral	43 [29.66]
Epidural	2 [1.38]
Subdural	30 [20.69]
Perimesenchephalic	4 [2.76]
Secondary Injuries Ischaemia/ Oedema	69 [47.56]
Pneumocephalus	60 [41.38]
Mass effect	50 [34.48]
Hygroma	21 [14.48]
Segmentation Types	37
Manual	
Interactive	108

Table 1: Clinical characteristics of study patients.

Manual and semi-automatic segmentation

A manual segmentation, in the present context, is the process where the users demarcate the areas of interest without any software assistance that will permit task completion at a faster pace. Instead, the users rely solely on the very fundamental tools of a segmentation software, which is typically a pen or a paintbrush that allows drawing upon the areas of interest. In a semi-automatic segmentation, smart brushes or other algorithms ease the segmentation process, enabling users to minimize their drawing input, which it turn reduces the time needed to complete the task. In contrast, in a fully automated segmentation method, the user lets the algorithm complete the segmentation task independently and corrects, when needed, the end results.

NCCTs with spontaneous ICH were segmented both manually and in a semi-automatic/interactive fashion. The open-source software ITKsnap was used to manually segment 37 volumes using the paintbrush mode and a round brush shape. The open-source software 3DSlicer was used to create semi-automated segmentations of 108 volumes. These segmentations were used as the gold standard to evaluate the paper's proposed algorithm.

The reference segmentations were generated by a single trained medical image analyst (AT) and later reviewed and corrected by a study neurosurgeon (MK) with 17 years' experience. Both raters were present during the review process, and all corrections to the segmentations were done after mutual agreement on the radiological findings. Each segmentation consisted of a binary volume mask with the same dimensions and coordinate system as the NCCT. Two discrete values, one and zero, of the binary volume mask represented the voxels classified as blood and background, respectively. The segmentation strategy, which was agreed beforehand, is summarized as follows: 1) Segmentations were conservative and only areas that the raters considered certain were marked as representing blood. Ambiguous areas with moderate or high uncertainty of being blood were considered as background (zero value); 2) The blood was only drawn onto the axial plane since the off-plane resolution was often low and therefore not informative; 3) A blood cluster had to be present in at least two consecutive slices in order to be marked as such in the segmentation process. An exception to this rule could occur only if the presence of blood on a patch of a given slice was certain. Metrics of the segmentations are depicted in Table 2.

Metric	Count/Mean (std)
Total Size	
Voxels	27079.73 (45550.07)
<min max>	<116 361419>
Volume (mm3)	17377.32 (23721.26)
<min max>	<58.51 141813.77>
Total surface extent (voxels)	
In Plane X	203.73 (76.99)
<min max>	<13 333>
In Plane Y	221.92 (85.27)
<min max>	<13 371>
Off Plane Z	18.59 (8.97)
<min max>	<2 56>
Clusters	88.66 (114.90)
<min max>	<1 729>
Largest cluster size (voxels)	19722.93 (42096.56)
<min max>	<68 356939 >
Largest cluster size (mm3)	12182.25 (20518.52)
<min max>	<34.30 140055.90>

Table 2: Haemorrhage characteristics of study patients.

Fully automated segmentation algorithm

The algorithm is a combination of a 2D and 3D intensity thresholding, morphological operations, region growing, and 2D entropy, which results in five segmented classes: brain, Cerebrospinal Fluid (CSF), blood, calcifications, and background.

The background class also includes noise emanating from surgical or endovascular materials (coils, clips, titanium plates, etc.). The class representing blood is the only class that is further filtered with a 2D entropy kernel. It is a fully automated software, which we have open-sourced, with its flowchart depicted in Figure 1. Briefly, it begins with the removal of irrelevant voxels outside the scalp and is followed by a brain extraction, where the skull and scalp are discarded.

Segmentation of the different classes inside the brain tissue is then achieved in two consecutive steps, with the blood class being further post-processed with the 2D entropy filter to produce the final blood segmentation result.

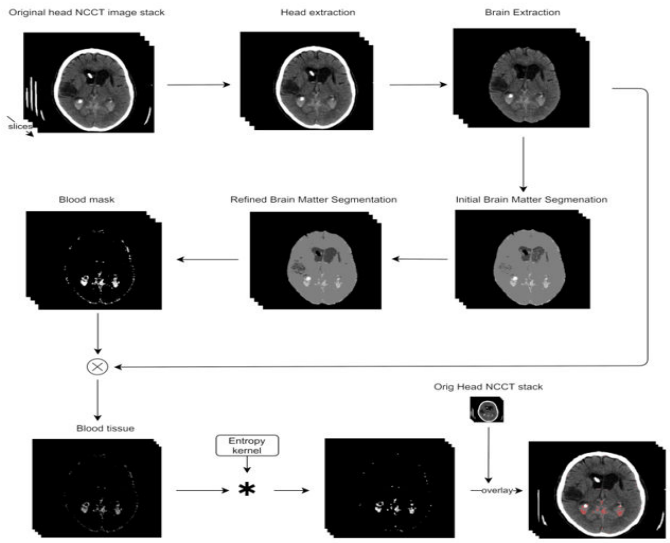

Figure 1: Algorithm flowchart of brain extraction.

The detailed algorithm steps are described as follows:

Head extraction, which in the present context means removal of irrelevant voxels outside the scalp (e.g. headrest), is based on the logarithmic histogram of the CT image. An HU bin range of [-1024, HUmax] was used, where HUmax was the maximum HU value of the image, and a bin width of 3 was used. Any HU values smaller than -1024 were clamped to this value. The local minima of the logarithmic histogram in the HU window [-750, -150] were then detected by finding the corresponding zero crossings of its first derivative approximated by finite differences with a step length of 3. If several minima were found, the one with the smallest bin counts was taken. If no minima were detected, the minimum bin count in the HU window was found and this bin was taken as the local minimum. The established local minimum was used as the HU threshold for separating the tissue, and a binary mask was created by simple thresholding. Using a 6-connected neighbourhood, the largest connected component in the head mask was found to represent the head. A morphological dilation using a disk-shaped structure element with a radius of 2 pixels was performed to compensate for the hard thresholding, and any cavities (holes) were filled in the axial slices of the 3D mask.

Brain extraction, which in the present context means the removal of the skull and scalp, used the extracted head mask as a starting point. Another logarithmic histogram was computed of the head-masked portion of the image using the same parameters as for the entire image, and the global maximum of the histogram was taken to represent the brain matter peak. An HU window of ± 60 units was placed symmetrically around the peak, and a crude brain mask was established by simple thresholding with the window limits. Again, the largest 6-connected component was taken, and the corresponding mask was stored as the initial brain extraction mask Mbrain,0.

To separate the brain from the skull and scalp, a coarse morphological erosion was first performed using a disk-shaped structure element with a radius of 7.5 mm. Then, the largest remaining 6-connected component was taken, and this component was dilated back using the same structure element, which effectively combines morphological opening with taking the largest connected component, with the order of the operations maximizing the separability of the structures. The resulting brain mask Mbrain,1 was then subtracted from the original brain mask Mbrain,0 established above in order to find the structures that were removed in this operation. Then, the smallest (max 25% of the largest component) connected components of these removed structures Mbrain,0 \ Mbrain,1 were restored to Mbrain,1. Next, a second morphological opening combined with an intermediate largest connected component thresholding was performed using a smaller disk-shaped structure element (radius 1.5 mm). Any cavities smaller than 1 ml were then filled in the axial planes.

To finish, a slight morphological erosion using a disk-shaped structure element with a radius of 3 pixels was performed, and the result was taken as the brain extraction mask.

Initial brain matter segmentation is a process where the brain matter is segmented into different classes by using the created brain mask (see "Brain extraction" above). The minimum and maximum HU values in each axial slice were determined, while the HU cut-off thresholds were determined by sorting the slice-wise HU limits in ascending order and taking the lower cut-off threshold at 30% of the sorted minimum values and the higher cut-off threshold at 50% of the sorted maximum values. These HU cut-off thresholds were used in the histogram of the brain tissue (3D volume of all axial slices). The zero crossings of the histogram's first derivative were calculated to detect the local maxima, while the zero crossings of the histogram's second derivative were used to detect the minima of the absolute value of the first derivative. Both the first and second derivatives were approximated using finite differences (step length 2). The highest local maximum was taken to represent the brain matter peak. If a secondary local maximum was detected before the brain matter peak, this represented CSF. Otherwise, the shallowest saddle point on the rising edge of the brain matter peak was assigned to CSF. Then, the shallowest saddle point on the falling edge of the brain matter peak was identified, and this HU value represented potential blood. If no local maxima and saddle points were detected around the brain matter peak, a constant HU offset of 25 was used as a threshold for probable CSF and blood. For differentiating CSF from the brain matter threshold, the local histogram minima were similarly found between the brain matter peak and the probable CSF peak that may or may not have been visible as a local histogram maximum. Of these potential minima, the one corresponding to the smallest histogram counts was treated as the threshold. When tried to differentiate the brain matter from the probable blood, no intermediate minima were typically detected, and a simple weighted average of the tentative peak HU values was used as a HU threshold, with 40% weight given to the brain matter peak and 60% weight to the peak for the probable blood. The selected HU thresholds were then used for an initial segmentation of the brain mask into CSF, brain matter, and potential blood.

Refined brain matter segmentation means the initial brain segmentation was next refined by varying the CSF-to-brain matter and the brain-matter-to-probable-blood thresholds. These variations were done using the HU windows [HUcutoff, min,

HUBM-20] for CSF and [HUBM, HUcutoff, max-15] for the probable blood, with the cut-off limits reported above. Using each threshold, the corresponding mean values and standard deviations of the connected CSFand probable blood component volumes were computed. For CSF, the threshold values used were either the minimum of the mean volume of the connected components or their standard deviations. This approach was used to result in a segmentation that should consist of connected components, which vary minimally in size but at the same time are not too large. The one corresponding to a smaller HU value was then considered the refined CSF-to-brain-matter threshold. A similar approach was used for the brain-matter-to-probable-blood threshold. As the next step, a simple bias compensation was performed to compensate for any partial volume effect (typically near the vertex, i.e. top of the skull) or CT reconstruction artefacts. The histograms of the axial slices were computed and the highest local maximum found. For each slice, the highest local maximum was assumed to correspond to the brain matter, and the peak shift with respect to the 3D brain matter peak represented the HU bias. The allowed peak shift was, however, limited to 5 HU values of the brain HU cut-off thresholds. Finally, an iterative refinement based on a region growing and a simple Gaussian fit was performed for the segmentation classes. Prior to the iteration, the intra-class mean HU values and standard deviations were computed, and the final HU cut-off limits were defined as μHU ± C•σHU. First, the brain matter class was grown using C=275. In each iteration, the class mask was dilated with a disk-shaped structure element, and any voxels falling outside the HU cut-off range were excluded. Then, the change in the relative class volume was computed and the iteration discontinued when the change between consecutive iterations fell below 5×10^{-6}. Second, the CSF class was dilated using C=1.5. Third, any values falling outside the high HU cut-off threshold (typically calcifications or metal objects) were dilated using C=0.25, and, finally, the probable blood was dilated using C=1.5. The tissue that had been classified as blood passed through an entropy filtering (slice-wise) using a disk structural element (size 2) and low thresholding of 1 to reduce the number of false-positive voxels. The sensitivity of the final segmentation in terms of false negatives *versus* false positives was ultimately determined by the magnitudes of C for the different tissue classes. For a manual refinement of the resulting segmentation, it was typically faster to erase unwanted regions, and larger values for C (more false positives) may have become favoured.

An optional post-processing step for the removal of the brain mask's outline could be used to reduce the false-positives in the brain's periphery. This outline was created with a morphological gradient using a 2D square 3 × 3 voxel kernel.

Algorithm evaluation

The fully automated algorithm was evaluated against the gold standard segmentations. Another much simpler method was also employed, and it used the brain-extracted tissue of the automated algorithm and thresholded it at different HU ranges to detect the blood. The HU range that created the most accurate blood segmentations (with respect to the gold standard ones) was then chosen to be compared with the fully automated algorithm. This approach is illustrated as a flowchart in Figure 2.

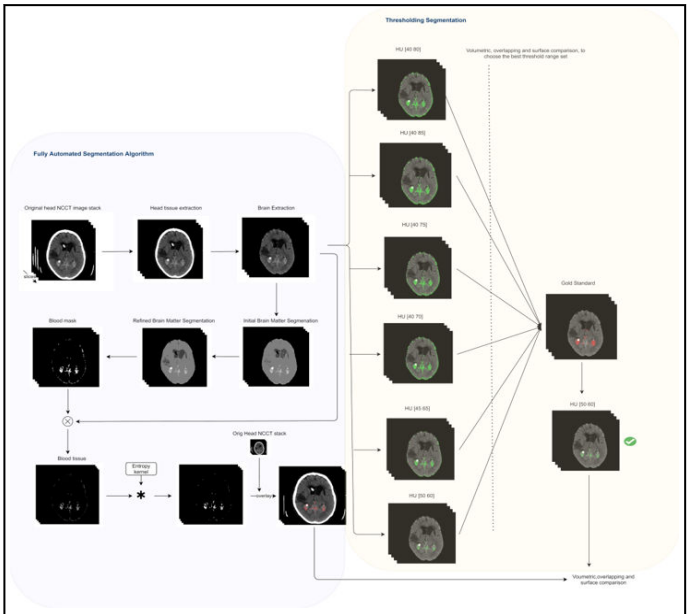

Figure 2: Algorithm evaluation flowchart.

Three types of similarity measures were used to assess the quality of the blood segmentation created by the algorithm: one that measures the blood volume (target-reference pair) (as depicted in Figure 3) segmentation, a second one that measures the volumetric overlap of the blood clusters (target-reference pair) (as depicted in Figure 4), and a third one that measures the surface extent occupied by blood pixels (as depicted in Figure 5). Such measures are well established in the literature [9]. More thorough quality metrics are available for both the patient and control groups in the supplementary material.

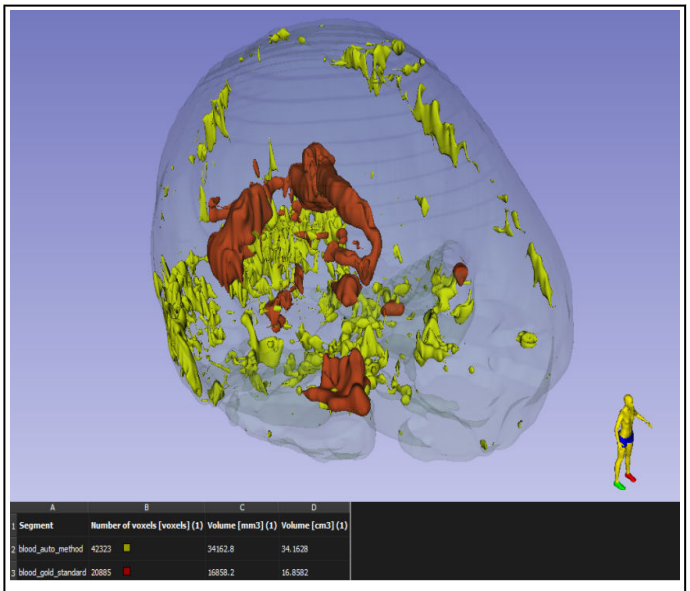

Figure 3: Volumetry of blood.

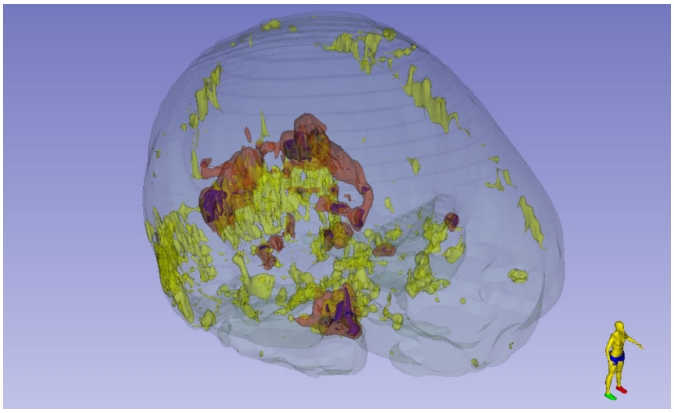

Figure 4: Volumetric overlap of the blood cluster.

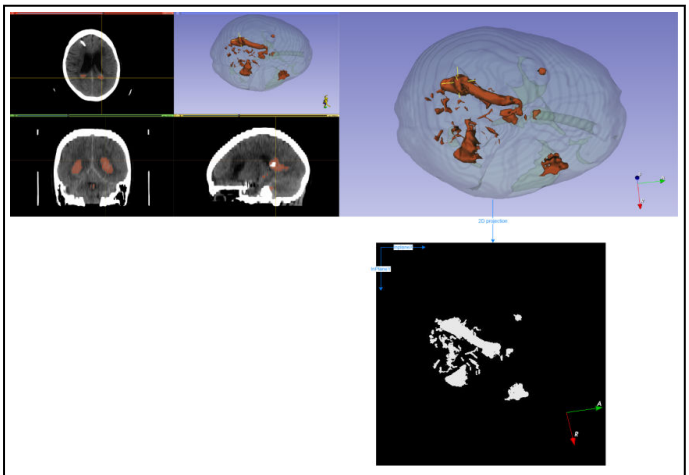

Figure 5: Surface metrics of blood pixels.

Patient and control groups: The fully automated algorithm's performance was compared with that of the simple intensity thresholding at typical HU ranges of blood. More specifically, the brain-extracted tissue was thresholded at different HU intensity ranges: (50,60), (45,65), (40,70), (40,75), (40,80), and (40,85). The same metrics that were used to evaluate the fully automated algorithm were also used to select the best HU range of the simpler thresholding approach. For the patient group (with intracranial haemorrhage), Jaccard and Dice metrics were calculated (since gold standard segmentations existed). For the control group, only the False Positives (FPs) were calculated (gold standard segmentations are not possible to do for healthy subjects).

The best simple intensity thresholding set (out of (50,60),(45,65), (40,70), (40,75), (40,80), and (40,85)) was selected to be compared with the fully automated algorithm. The best set was the one producing the best Jaccard coefficient. All different pairs were compared (i.e. Jaccardset1 *vs.* Jaccardset2) and their differences were examined using a Wilcoxon signed-rank test since their distribution was not normal (the Shapiro test was used to evaluate normality). Comparisons with the algorithm were done for every metric pair (i.e. Jaccardalgorithm *vs.* Jaccardthr50-60, Dicealgorithm *vs.* Dicethr50-60, TPsalgorithm *vs.* TPsthr50-60 and so on).

The statistical difference between the pairs was examined using a Wilcoxon signed-rank test since the paired samples did not follow a normal distribution. All statistical analyses were conducted using Python's statistical module stats of the SciPy library.

RESULTS

Segmentation times

The mean time required to complete a manual segmentation (using ITKsnap) of one head CT scan with intracranial haemorrhage (88% were SAH cases) was 144.4 min (standard deviation 134.1 min). The mean time required for the review and corrections by the second rater was 10.6 min per manually segmented case. On average, every patient's head CT scan (volume) consisted of 38.8 reconstructed (MPR) axial slices.

The mean time required to complete an interactive segmentation (using 3DSlicer) of one head CT scan with intracranial haemorrhage was 24.7 min (standard deviation 24.7 min). The average time required for the review and corrections by the second rater was 8 min per interactively segmented case. On average, every patient's head CT scan (volume) consisted of 36.4 reconstructed (MPR) axial slices.

Performance of algorithms

Threshold algorithm: Of all threshold sets, the one with intensity range of (50,60) provided the best results. The set (45,65) provided very similar results to the set (50,60) (p_{value}= 0.1, Wilcoxon test). The set (50,60) was chosen (Jaccard (mean,std)=(0.08,0.08)) because of the slightly smaller standard deviation than with (45,65) (Jaccard (mean,std) = (0.08,0.09)).

Fully automated algorithm-patient cases: The ICH segmentation results of the quality assessment of the fully automated segmentation algorithm are presented in Table 3. The performance of the fully automated algorithm was evaluated using the ICH patient group images, which had been both manually and interactively segmented as described above. In the volumetric analyses, the signed relative voxel difference with respect to the gold standard was 2461.2, which was similar to the value of 2513 for the simple thresholding method. Both the fully automated and the simple thresholding method oversegmented blood in a similar way (p_{value}=0.37, Wilcoxon signed-rank). However, this oversegmentation was less scattered in the case of the fully automated method, and the simple thresholding method resulted in a much larger number (p_{value} << 0.05) of small-sized clusters. The smaller size of the clusters can be deducted from the total number of detected voxels-to-clusters ratio, which was 70.81 (voxels/clusters) for the fully automated algorithm and 17.80 for the simple thresholding method (at HU(50 60)). In the overlap analysis, the Jaccard ((p_{value} << 0.05)) and Dice (p_{value} << 0.05) were higher for the fully automated algorithm, suggesting a better overall performance. The fully automated algorithm segmented more true-positive voxels than the simple thresholding method (p_{value} << 0.05), whereas such a difference was not evident for the false positives (p_{value}=0.16). A typical performance (Dice=0.20) of the fully automated algorithm is illustrated in Figure 6. A performance well above the average (Dice=0.59) is shown in a supplementary file. Sixty-three (43%) of the 145 ICH cases had a Dice coefficient of less

than 0.1, while 56 (38.6%) had values above 0.2. Out of the 145 ICH patients' head CT scans (volumes), the fully automated algorithm did not detect a single true-positive voxel in two cases and detected only one voxel in another case. Therefore, at patient level the fully automated method missed 2 out of 145 patients, while the simple thresholding missed none. At slice level, the fully automated method segmented correctly (at least one correct voxel per slice) 92.98% of the positive slices (2277 out of 2449 positive slices). At voxel level, the fully automated method segmented correctly (true-positive rate) 54.12% and incorrectly (false-positive rate) 0.63% of the positive voxels. Similarly, the simple thresholding method segmented correctly 96% of the slices (2351 out of 2449), while at voxel level it segmented correctly (true-positive rate) 31.75% and incorrectly (false-positive rate) 0.62% of the positive voxels.

Metric	Count/Mean (std)
Total Size	
Voxels	76070.61(48084.62)
<min max>	<9510 242860>
Volume (mm3)	52629.01(30888.87)
<min max>	<7641.04 167191.71>
Clusters	1074.26(429.16)
<min max>	<219 2967>
Largest cluster size (voxels)	37715.05(39809.65)
<min max>	<1999 180329 >
Largest cluster size (mm3)	25583.63(25308.33)
<min max>	<1551.07 117773.08>
Signed relative voxel difference	2461.23(5055.83)
<min max>	<-88.30 36904.31>
Total surface extent (voxels)	
In Plane X	291.02(18.71)
<min max>	<240 384>
In Plane Y	370.50(20.78)
<min max>	<307 426>
Off Plane Z	33.07(6.87)
<min max>	<17 65>
Dice	0.20(0.20)
<min max>	<0 0.73>
Jaccard	0.12(0.14)
<min max>	<0 0.57>
True positives	14656.67(25510.00)
<min max>	<0 155921>
False positives	61413.94(40453.15)
<min max>	<779 214077>

Table 3: Quality assessment of the fully automated segmentation algorithm.

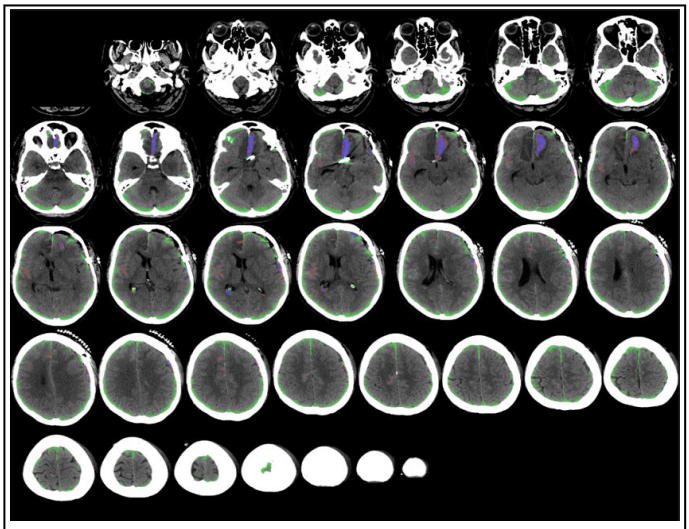

Figure 6: Average performance of the algorithm on an axial postoperative NCCT of a patient with SAH, SDH, IVH, and ICH.

Fully automated algorithm - control cases: The results are depicted in Table 4. The algorithm's performance was compared with the simple intensity thresholding method (HU (50,60) provided the smallest number of false positives). The results of the simple thresholding method on the control images are summarized in the supplementary files. A more detailed description of the performance is reported also in supplementary files. In the volumetric analyses, the false positive voxels detected by the fully automated algorithm were fewer (p_{value} << 0.05) than those detected by the simple thresholding. The total number of detected voxels-to-clusters ratio was 32.81 (voxels/clusters) for the fully automated algorithm, indicating that the false positives were aggregated in fewer clusters of larger sizes than when using the simple thresholding method (voxels/clusters ratio of 9.68). One out of the 150 control cases could not be processed. An average performance of the fully automated algorithm (which led to the average amount of false

positives in the whole control group) is shown in Figure 7. Cases in which the algorithm produced the minimum and maximum amount of false positives are presented in supplementary files.

The fully automated algorithm as well as the simple thresholding segmented false positives in all control cases without acute bleedings (patient level).

At slice level, the fully automated method correctly recognized 7.57% of the negative slices (392 out of 5178) and incorrectly segmented blood in 4786 slices (out of 5178).

The simple thresholding method correctly recognized 7.33% of the negative slices (380 out of 5178) and incorrectly segmented blood in 4798 slices (out of 5178).

Metric	Count/Mean (std)
Total size	
Voxels (False Positives)	48637.91(27990.37)
<min max>	<15584 182469>
Volume (mm3)	39480.89(20087.41)
<min max>	<13512.27 129232.86>
Clusters	1482.17(686.95)
<min max>	<485 4723>
Largest cluster size (voxels)	18910.40(21671.83)
<min max>	<1318 138320 >
Largest cluster size (mm3)	15141.21(16457.88)
<min max>	<1263.66 109409.18>
Total surface extent (voxels)	
In Plane X	286.62(13.51)
<min max>	<245 330>
In Plane Y	359.99(19.50)
<min max>	<315 418>
Off Plane Z	32.15(5.95)
<min max>	<23 70>

Table 4: Quality assessment of the fully automated segmentation algorithm.

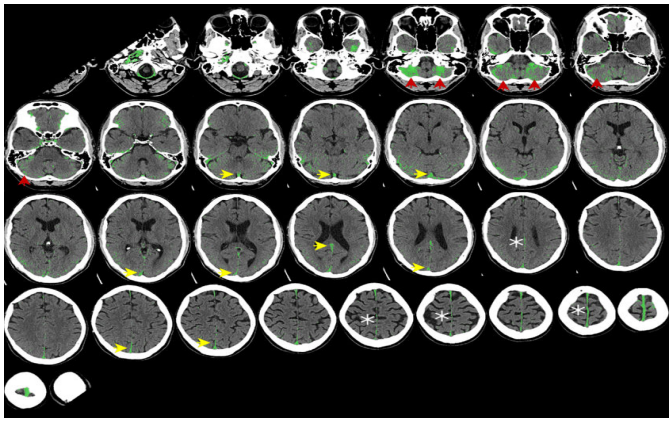

Figure 7: Average performance of the fully automated algorithm in segmenting a control case.

DISCUSSION

This paper presents the challenges encountered when segmenting NCCT images, and we open-source a fully automated segmentation algorithm that is likely to assist experts in segmenting intracranial bleedings, particularly on images with high noise content.

The fully automated algorithm was developed using a hybrid method that combines histogram information, region growing, morphological operations, and entropy. The algorithm correctly segmented blood in 143 out of 145 patients. Its average performance using the Dice coefficient metric was 0.20, which we consider a moderate value. The performance of published algorithms using the Dice metric ranged from 0.6 to 0.92, but, as discussed later, it reflects the segmentation quality of carefully selected artefact-free images. The Dice coefficient is among the most common metrics used to assess segmentation quality. Dice punishes more a mismatch of small clusters than a mismatch of larger clusters. This means that in SAH images the blood is often widely distributed and present in relatively small voxel clusters, and therefore, Dice coefficients for the mismatch (with respect to the gold standard) are significantly lower than for other types of haemorrhages (such as ICH, EDH, and SDH, which are more contained and larger in volume). This phenomenon is further exaggerated in cases of diffuse SAHs, where multiple very small clusters of blood are widely dispersed throughout the subarchnoid space. In fact, this can be observed in the work of Boers et al. [14], which presents a Dice coefficient of 0.64 ± 0.20 (inter-rater variability across two experts) for segmenting SAH blood in head CT scans. Our algorithm failed to segment any blood in three SAH cases. These three cases contained not only major image artefacts but also represented very challenging SAH cases (very small cluster remnants of blood in subarachnoid spaces and ventricles). We believe that if our algorithm, which is not meant for diagnostics, misses around 2% of SAH cases with a very small amount of blood, the effect on the overall segmentation time of tens or hundreds of SAH cases is insignificant. In other words, these few cases can

perhaps be manually segmented, especially since a manual segmentation of small bleeds is not an overly cumbersome task. Removing a thin layer of the brain's cortex will improve the results, especially when other bleeding types, such as SDH and EDH, are not present. The average segmentation time for a single NCCT with intracranial blood was about 3 min on a standard laptop. Furthermore, as the algorithm performed relatively well even when it was used for segmenting images with major artefacts, we believe that this segmentation tool, which functions in a simple laptop, assists this crucial step in developing computer vision algorithms for clinical use.

Many of the previous algorithms published have been built and validated using highly selected datasets. Such image selection does not necessarily represent the high variability of the images encountered in a clinical setting. In contrast to these algorithms, we did not develop the fully automated algorithm to detect blood in highly selected images, but to assist in creating new clinical algorithms based on highly variable and artefact-rich datasets. We did not exclude any images based on their quality or on the medical interventions that the study patients had undergone. Moreover, we used both preoperative and early postoperative images in order to increase the complexity and variability of the dataset. Overall, 72% of the cases were acquired after medical interventions, and therefore, the images may have included, for example, air, aneurysm clips, aneurysm coils, Onyx embolization material, ventriculostomies, and multiple subtypes of bleedings. Since we are not aware of any similar studies using postoperative or posttreatment images in the development of segmentation or computer vision algorithms, true performance comparisons are challenging to conduct. In general, algorithms built to fulfil the same objectives can be compared, upon the same benchmark data, using segmentation metrics like Dice and others.

Because comparisons with other complex datasets or algorithms were impossible to conduct, multiple metrics were computed for the fully automated algorithm in an effort to shed light on various aspects of segmentation quality. For the same reason, we opted to create a second, much simpler approach, namely a thresholding. This approach was used to form a simple baseline during the algorithm's development process. If the algorithm had performed similarly to the best possible thresholding, it would have been a clear indication of a useless result. The moderate performance of the algorithm (Dice coefficient 0.20) can be attributed to the following reasons. First, most (87%) of the patients had SAH, which will be more penalized, as explained above. Second, most (72%) of the patients had undergone an intervention (open surgery or endovascular treatment), which had introduced major artefacts and image distortions. Such artefacts and distortions introduce false positives, while also prohibiting the use of powerful techniques, such as brain atlas priors, which have the potential to increase segmentation quality. Third, our method of creating the gold standard segmentations may have led to undersegmentation of blood. In other words, when a cluster of blood was not visible in at least two consecutive slices or when there was an ambiguity about the borders of the blood clusters, we did not segment the corresponding voxels as blood. Therefore, the gold standard segmentation may be undersegmented, and the "false positives" segmented by the fully automated algorithm may, in many cases, be true positives. Therefore, our pixel-level results are likely better than reported. Finally, the use of the images' intensity distribution (histogram) to segment structures has inherent limitations. Such an approach uses only the intensity content of the image, while the spatial content remains unutilized. In summary, considering the highly heterogeneous images in this study, we believe it is not perhaps appropriate or relevant to expect a high segmentation accuracy at voxel level. The algorithm performs better with artefact-free images; out of the 10% of patients whose Dice coefficient was larger than 0.5, about 60% had preoperative images. Overall, the algorithm segmented 54.12% of the gold standard voxels correctly (true-positive rate), but falsely (some of these may have been true positives, as explained above), and 0.63% of the negative voxels in the gold standard images (false-positive rate). We also assessed the performance of the fully automated algorithm in other than the intended context, i.e. in segmenting control cases without acute bleedings. Expectedly, the fully automated algorithm segmented false positives in all control cases, indicating that the algorithm is not suitable for clinical diagnostics but for segmenting challenging head CT images with acute bleeding. We consider this rate of false positive segmenting as a safety measure against a non-intended use.

A high-throughput segmentation process of medical images is one of the most important steps for a successful training of diagnostic machine learning models. When creating a dataset for training, image variability is believed to increase the likelihood of a successful training of a clinically applicable algorithm. For a clinical expert needed to segment such a training material, a fully automated segmentation method that will assist in completing the task faster is probably helpful. The requirements for an assistive fully automated segmentation algorithm are significantly different than for a diagnostic algorithm, as the main objective for the former is robustness on a wide variety of images and for the latter utmost segmentation accuracy. The created segmentation algorithm is intended to be helpful not only for a diverse quality of preoperative images, but also for postoperative images with multiple artefacts since such images are surely useful in training diagnostic algorithms for clinical use.

These image processing steps were chosen because, regardless of their limitations, they are less likely to lead to failure in segmenting images that exhibit high variability and increased artefact content.

Classical (non-machine learning) image processing approaches, like the one presented here, have the advantage of being fast and relatively simple. Moreover, the presented algorithm can be used in everyone's laptop or desktop computer and does not require high computing power. In addition, to our knowledge, such a detailed description of the challenges that should be tackled for a successful segmentation of intracranial bleedings has not been reported before.

The study also has some shortcomings. The datasets used cannot reflect the vast variability that exists in medical centres globally. However, the image variability is still rather extensive.

Moreover, we decided to open-source the 145 patient cases along with their gold standard segmentations in order to establish a publicly open complex dataset, which can be widely used to standardize reporting of algorithm performances. In fact, our dataset enables performance reporting not only at patient level, but also at slice and voxel levels. To our knowledge, this is the first such open-sourced dataset. Another shortcoming is that the developed segmentation algorithm does not produce very high segmentation accuracy. If the accuracy were to be very high, the algorithm would not serve as a segmentation algorithm, but rather as a diagnostic algorithm. Nevertheless, it will be able to perform fairly well in segmenting very challenging images, which are needed to create truly clinically useful diagnostic algorithms.

CONCLUSION

The developed fully automated segmentation algorithm can assist clinical and even non-clinical experts in creating training material for machine learning algorithms that can detect SAH in head CT scans. In order to transparently and openly advance brain haemorrhage-related algorithm development, the computed python algorithm is available in GitHub. Moreover, in an attempt to help standardize the assessments of algorithm performances, we have openly shared the segmented dataset of preoperative and postoperative images. We believe that the shared algorithm and dataset can serve as benchmark material for similar studies.

ACKNOWLEDGMENTS

The authors thank Miika Leminen and Taru Hermens for providing administrative support and for partly reviewing the manuscript. We also thank Heikki Peura and Jenni Wennervirta for reviewing the manuscript and radiological information. Finally, Eero Salli is thanked for valuable advice and suggestions in the processes of fetching the study images from the PACS, for image anonymization, and for reviewing and commenting on the manuscript. The current work was conducted as a part of the CleverHealth network ecosystem and the project AI head analysis.

REFERENCES

1. Raj R, Bendel S, Reinikainen M. Costs, outcome and cost-effectiveness of neurocritical care: a multi-center observational study. Critical Care. 2018;22(1): 1-10.
2. An SJ, Kim TJ, Yoon BW. Epidemiology, risk factors, and clinical features of intracerebral hemorrhage: an update. Journal of stroke. 2017;19(1): 157-192.
3. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ digital medicine. 2018;1(1): 1-7.
4. Smith-Bindman R, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. Health affairs. 2008;27(6): 1491-1502.
5. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal. 2017;2017: 359.
6. Nishie A, Kakihara D, Nojo T. Current radiologist workload and the shortages in Japan: how many full-time radiologists are required? Japanese journal of radiology. 2015;33(5): 266-272.
7. Rosenkrantz AB, Hughes DR, Duszak Jr R. The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology. 2016;279(1): 175-184.
8. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging. 2015;15(1): 1-28.
9. Heimann T, Van Ginneken B, Styner MA. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE transactions on medical imaging. 2009;28(8): 1251-1265.
10. Choi G, Park DH, Kang SH, Chung YG. Glioma mimicking a hypertensive intracerebral hemorrhage. Journal of Korean Neurosurgical Society. 2013;54(2): 125.
11. Grech R, Grech S, Mizzi A. Intracranial calcifications: a pictorial review. The neuroradiology journal. 2012;25(4): 427-451.
12. Berberat J, Grobholz R, Boxheimer L, Rogers S, Remonda L. Differentiation between calcification and hemorrhage in brain tumors using susceptibility-weighted imaging: a pilot study. American Journal of Roentgenology. 2014;202(4): 847-850.
13. Kiroglu Y, Calli C, Karabulut N, Oncel C. Intracranial calcifications on CT. Diagnostic and Interventional Radiology. 2010;16(4): 263.
14. Boers AM, Zijlstra IA, Gathier CS. Automatic quantification of subarachnoid hemorrhage on noncontrast CT. American journal of neuroradiology. 2014;35(12): 2279-2286.