Expert-Level Detection of Acute Intracranial Hemorrhage on Head Computed Tomography using Deep Learning

Wei-Cheng Kuo
Christian Haene
Pratik Mukherjee
Esther Yuh
Jitendra Malik

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-13
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-13.html

May 1, 2021

Acknowledgement

This study was supported in part by California Initiative to Advance Precision Medicine (California Governor's Office of Planning and Research). Christian Häne also received funding from the Swiss National Science Foundation (Early Postdoc.Mobility Fellowship #165245). We are also grateful to Amazon Web Services (AWS) who provided compute time.

Expert-Level Detection of Acute Intracranial Hemorrhage on Head Computed Tomography using Deep Learning

by

Wei-Cheng Kuo

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jitendra Malik, Chair
Professor Bin Yu
Professor Alexei Efros

Spring 2019

Expert-Level Detection of Acute Intracranial Hemorrhage on Head Computed Tomography using Deep Learning

Abstract

Expert-Level Detection of Acute Intracranial Hemorrhage on Head Computed Tomography using Deep Learning

by

Wei-Cheng Kuo

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Jitendra Malik, Chair

Computed tomography (CT) of the head is used worldwide to diagnose neurologic emergencies. However, expertise is required to interpret these scans, and even highly trained experts may miss subtle life-threatening findings. For head CT, a unique challenge is to identify, with perfect or near-perfect sensitivity and very high specificity, often small subtle abnormalities on a multislice cross-sectional (3D) imaging modality that is characterized by poor soft tissue contrast, low signal-to-noise using current low-radiation-dose protocols, and a high incidence of artifacts.

We view the task as a semantic segmentation problem and tackle it with a patch-based fully convolutional network (PatchFCN). To develop the model, we collected a dataset of 4396 head CT scans performed at University of California at San Francisco and affiliated hospitals, and compared the algorithms performance to that of 4 American Board of Radiology (ABR) certified radiologists on an independent test set of 200 randomly selected head CT scans. Our algorithm demonstrates the highest accuracy to date for this clinical application, with a receiver operating characteristic (ROC) area under the curve (AUC) of 0.991 0.006 for identification of exams positive for acute intracranial hemorrhage, and also exceeded the performance of 2 of 4 radiologists. We demonstrate an end-to-end network that performs joint classification and segmentation with exam-level classification comparable to experts, in addition to robust localization of abnormalities including some that are missed by radiologists, both of which are critically important elements for this application. Furthermore, we demonstrate promising multiclass segmentation and detection results competitive with the state-of-the-art in an exploratory study.

Finally, we study how to scale up the data without naive labeling by building a cost-sensitive active learning system. Our method compares favorably with the state-of-the-art, while running faster and using less memory. The approach is inspired by observing that the labeling time could vary greatly across examples, we model the labeling time and optimize the return on investment. We validate this idea by core-set selection and by collecting new

data from the wild. Our method shows good estimation of human annotation time and clear performance gain under fixed annotation budget.

To Yen-Ming and Hui-Hsi

Contents

Contents **ii**

1 Introduction **1**

2 PatchFCN **3**
2.1 Introduction 3
2.2 Method 4
2.3 Experiments 6
2.4 Summary 11

3 Expert-level PatchFCN **12**
3.1 Introduction 12
3.2 Method 12
3.3 Evaluation and Benchmark with Radiologists 16

4 Cost-Sensitive Active Learning **25**
4.1 Introduction 25
4.2 Cost-sensitive Active Learning 27
4.3 Data Collection 28
4.4 Experiments 29
4.5 Conclusion 32

5 Conclusion **33**

Bibliography **34**

Acknowledgments

It's with much gratitude that I'm writing this thesis. There are so many wonderful people helping me along the way. Without them, this would never have been possible.

First and foremost, I'd like to thank my advisor Jitendra Malik for his steadfast, unwavering support throughout my entire PhD. There was a period of time when the things I worked on just did not work. Jitendra was very patient and encouraging, telling me that it's normal that things don't work sometime, and that we should fail fast if we are to fail. Then we changed direction and things began to work again. In addition, I really appreciate that Jitendra supports me to work on biomedical imaging project as the main theme of my PhD. To my knowledge, I'm the only person doing this among the computer vision group at Berkeley, but Jitendra doesn't discount the importance of the problem because nobody else is doing it here. He said "Everyone has his/her way of success. When things start working for you, you need to double down on it." Because of his belief and support, we're able to produce good results at the end.

In addition, I'd like to thank my collaborators Dr. Esther Yuh and Dr. Pratik Mukherjee. Thank you for being so available to discuss things anytime of the day, including weekends. I'd never forget those weekends that we work late in the UCSF office to figure things out. Esther and Pratik's clinical expertise shapes the direction of the project and guide the technical approach we take. This thesis really wouldn't happen without you. On top of that, I'd like to thank my collaborator Christian Haene. Thank you for working closely with me for two years from start to finish, even through the rejections of our papers. Your insight has great influence on the project. The patch idea would not have come into being without you. I especially appreciate your attention to both the big picture and details – all the places where things can go wrong.

Finally, I want to give thanks to many post docs, alumni, and students from the Berkeley computer vision group. Thank you Bharath for being my first mentor at Berkeley. It's been 4 years since we finish the DeepBox project, but I still vividly remember how you patiently train me through my rookie year. These are things I've never hoped for and things I'd never forget. Thank you David for letting me work with you on the video dataset. We tried many things and I learn a lot from you about how to make a dataset useful to the community. Thank you Katerina for guiding me through the tracking project. Although we don't have a paper, I did learn a lot from you about video analysis and just research in general. Thank you Saurabh, Shubham, Georgia, Pulkit, Philip, Evan, Jeff, Judy, Abhishek for always being there to answer my questions, many of which are not very smart. You guys really set great patterns of productive researchers for me – always asking good/insightful questions and rigorously studying them through control experiments. Thank you Ke and Deepak for being my companions through the PhD. We took so many classes together in the first few years. I'd always cherish the good times and debates we had doing homeworks, projects and even taking exams together.

Apart from people in the lab, I'd also give a special shoutout to the friends/families in the church in Berkeley that support me through these years. You guys make the long PhD

possible and enjoyable. Although I spent a lot of screen time, I'd always remember the time we cook, eat, sing, pray, hike, play frisbee or road trip together. You guys make my home away from home, and I could not thank you enough for that. And last but not least, thank you Emily for everything. It's been a tremendous blessing to have you with me.

Chapter 1

Introduction

As a result of falls, collisions, or accidents, traumatic brain injury (TBI) is the leading cause of accident-related deaths and disability worldwide (more than 30%), and 153 people die from TBI-related injuries each day in the US [34]. It is estimated that 5.3 million people in the US are living with TBI-related disability [27]. In emergency departments (EDs), head computed tomography (CT) scans are routinely performed on patients under evaluation for suspected TBI, based on history and other clinical criteria. Since the brain is vulnerable to irreversible injury within a matter of minutes, immediate diagnosis and treatment are essential. Well-trained neuroradiologists can reliably read these scans, but even among them, the agreement is imperfect [17]. In some EDs, the initial interpretation and decisions may be made by emergency physicians, whose reads are significantly less reliable than those of radiologists [7].

Evaluation for acute intracranial hemorrhage plays a critical and decisive role in the clinical management of these conditions. It is critical for deciding on the need and approach for emergent surgical intervention. It is also essential for allowing the safe administration of thrombolytic therapy in acute ischemic stroke. Since time is brain, a computer vision system that rapidly and reliably detects emergency TBI findings, such as acute intracranial bleeding, would potentially be a life-saving innovation, reducing both death and long-term disability. In practice, TBI is manifested on head CT by the presence of intracranial hemorrhage, which appears in a large variety of sizes and morphologies.

Advances in computer vision techniques such as deep learning have demonstrated tremendous potential for extracting clinically important information from medical images. Examples include grading of diabetic retinopathy on retinal fundus photographs [11], detection of metastases in histologic sections of lymph nodes [2], and classification of images of skin cancer [8], with accuracies comparable to or, in some cases, exceeding that of experts. In contrast to these applications, many radiological imaging studies, such as CT and magnetic resonance imaging (MRI), are cross-sectional, or three-dimensional, in nature and thus comprised of volumetric stacks of images rather than single images. The 3D nature of such exams presents an extra challenge. An additional unusual challenge regarding head CT is the need to identify, with perfect or near-perfect sensitivity, often tiny subtle abnormali-

ties occupying around 100 pixels on noisy, low-contrast images in a large 3D volume that comprises more than a million pixels. Finally, although perfect sensitivity at exam-level classification is the most crucial goal, concurrent localization of abnormalities on head CT is also important, since physicians will always need to personally visualize and confirm the locations of abnormalities on a head CT exam, in order to judge the need and approach for surgical intervention.

We propose a simple yet effective technique called PatchFCN (Patch-based Fully Convolutional Networks). Inspired by the observation that radiologists primarily rely on local cues to make decision, PatchFCN finds a good tradeoff between batch diversity and the amount of context. Our analyses show that it is more effective than vanilla FCN for hemorrhage detection. Using a strong pixel-level supervision approach and a training data set of 4396 scans, PatchFCN demonstrates the highest classification accuracy to date compared to other deep learning approaches [36, 18, 4, 5, 25], and also concurrently localizes these abnormalities. We demonstrate that it identifies many abnormalities missed by experts, along with promising results for multiclass hemorrhage segmentation and detection while preserving two-class accuracy.

To scale up the model further, we propose a cost-sensitive active learning framework that operates by modeling the labeling time for each exam. The intuition is every exam comes at a different labeling cost, which needs to be taken into account for active learning to be effective in practice. Experiments show that our method achieves state-of-the-art results on core-set selection setting with less computing cost and work well on data in the wild as well.

Chapter 2

PatchFCN

2.1 Introduction

Deep learning techniques have been successful recently in detecting intracranial hemorrhages, e.g. 3D classification [1, 36] supervised by text reports, 2D classification [18], instance segmentation [4]. However, to our knowledge, no semantic segmentation approach has shown performance competitive with human experts at exam level or with other methods at pixel level.

We propose to solve the detection and segmentation problem *jointly* as a semantic segmentation task. Segmentation offers many advantages over classification, including better interpretability, and quantifiable metrics for disease prognosis [4, 18]. Our approach is fundamentally different from Mask R-CNN[12, 4], which treats hemorrhage segmentation as an instance segmentation problem, or detection of discrete objects. Since hemorrhage is fluid (stuff, e.g., water, sky, grass) [9] and takes on highly variable morphologies often without well-defined boundaries separating discrete objects (things, e.g. cup, car), semantic segmentation is a simple elegant approach with none of the requirements of object detection and region processing associated with Mask R-CNN.

Among existing pixel-wise labeling techniques, fully convolutional networks [20] (FCN) are successful and widely adopted for such tasks in computer vision [20] and the medical imaging community [42, 26]. Most computer vision practitioners use whole images as inputs for their FCNs following [20]. This is in contrast to how patch-based FCN training has been successful in applications such as retinopathy [42], MRI [26], and X-ray/CT imaging [38, 41]. Despite the wide adoption, there exists no systematic study on why patches improve FCN in many cases.

We propose PatchFCN and show that it outperforms standard FCN in localizing hemorrhages. Since no public dataset is available, one important challenge we face is to acquire pixelwise labeled data. Unlike the approaches that learn from text reports [1, 36], we collect

This chapter is based on the work PatchFCN [16] done with Christian Hne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Statements about past work should be read with this context in mind.

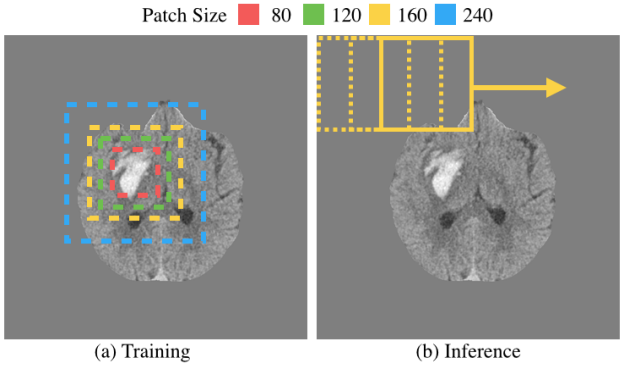

Figure 2.1: PatchFCN train on small patches and test in sliding window fashion. The colored boxes show different patch sizes in the context of a hemorrhage.

a dataset of 591 scans annotated *pixelwise* for the presence of hemorrhage by expert radiologists to validate PatchFCN. Using 100x smaller data, PatchFCN significantly outperforms weakly supervised methods [1, 36] on classification tasks.

We analyze the following factors to better understand the performance gains of PatchFCN: 1) batch diversity, 2) amount of context, and 3) sliding window inference. We find that PatchFCN outperforms FCN by finding an optimal trade-off between batch diversity and the amount of context. In addition, sliding window inference helps to bridge the gap of train/test time and consistently improve performance. We hope these findings would benefit other segmentation tasks where patch-based training is effective.

2.2 Method

The goals for hemorrhage detection are to find out: 1) whether a stack contains hemorrhage, and 2) where the hemorrhage is within the stack. In practice this may be used by the radiologists/neurosurgeons to assess the risk level of the patient and triage the patient to immediate surgical evacuation, monitoring in the intensive care unit (ICU), or routine monitoring on the hospital ward. Inspired by existing works [26, 42, 38, 41, 13], we propose to solve both tasks with PatchFCN as follows (see Fig.2.1):

Patch-based Training:

We train an FCN on random small patches cropped from the whole images centered on foreground. The model learns to predict the binary pixel label within the patches. For head CT data, the intuition of patch-based training comes from how radiologists make decisions – the morphology of contrast region is often a crucial cue for deciding whether it represents pathologies. Similarly, PatchFCN causes the network to make its decision based on the local image information without relying on excessive context. In addition, small patches allow larger batch size and hence higher batch diversity to stabilize network training. As most

convolutional networks have built-in batch normalization e.g. [40], PatchFCN leverages it by finding a good trade-off between large minibatch and adequate context for the task.

Patch-based Inference:

At test time, we evaluate the images in a sliding window fashion, as opposed to the typical fully convolutional inference. Sliding window at test time avoids any domain shift which occurs when training on small patches and evaluating fully convolutionally on the whole image. This is because the paddings present in convolution layers make a patch in the context of a whole image not the same as the patch by itself. Let the input image be of size H and the patch size C, then the total number of windows is given by $N = \lceil \frac{\beta H}{C} \rceil^2$, where $\beta > 1$ is an adjustable parameter for the window overlap. As multiple predictions are made for each pixel, we simply average their scores. The frame-level score is obtained by averaging the pixel scores within the frame. To get stack-level scores from pixel scores, we first take L^p-norm over the frame to obtain a stack-frame score. The stack score is defined as the maximum stack-frame score within a stack. p is treated as a hyper-parameter and tuned on the trainval set.

Data Collection:

Our dataset consists of 591 clinical head CT scans performed over 7 years from 2010 to 2017 on 4 different 64-detector-row CT scanners (GE Healthcare, Siemens) at our affiliated hospitals. We use the word "stack" for each patient's head CT scan, and the word "frame" for each individual slice of the stack. The scans were anonymized by removing patient-related meta-data, skull, scalp, and face. Board-certified senior neuroradiologists who specialize in TBI identified all areas of hemorrhage in our dataset. Our data contains the typical spectrum of technical limitations seen in clinical practice (e.g. motion artifact, "streak" artifact near the skull base or in the presence of metal), and also contains all of the subtypes of acute intracranial hemorrhage, including epidural hematoma, subdural hematoma, subarachnoid hemorrhage, hemorrhagic contusion, and intracerebral hemorrhage (see Fig. 2.2 for examples). We randomly split the data into a trainval/test set of 443/148 stacks for development and internal validation. The hyper-parameters of PatchFCN are tuned within the trainval set.

Implementation Details:

We choose a DRN-38 backbone because it performs competitively among many network designs [40]. Regarding the inputs, we clip the dynamic range of raw data at -40 and 90 Hounsfield unit (HU), and then rescale the intensity to lie within $[0, 255]$. Image size is 512×512. In both training and test time, we use a patch size of 240 unless stated otherwise. We utilize the z-axis context by fusing the adjacent frames with the center frame at the input (3 channels in total). The optimization is done by SGD with momentum following [40] setup. We train the network from scratch without using ImageNet pretraining, as we do not

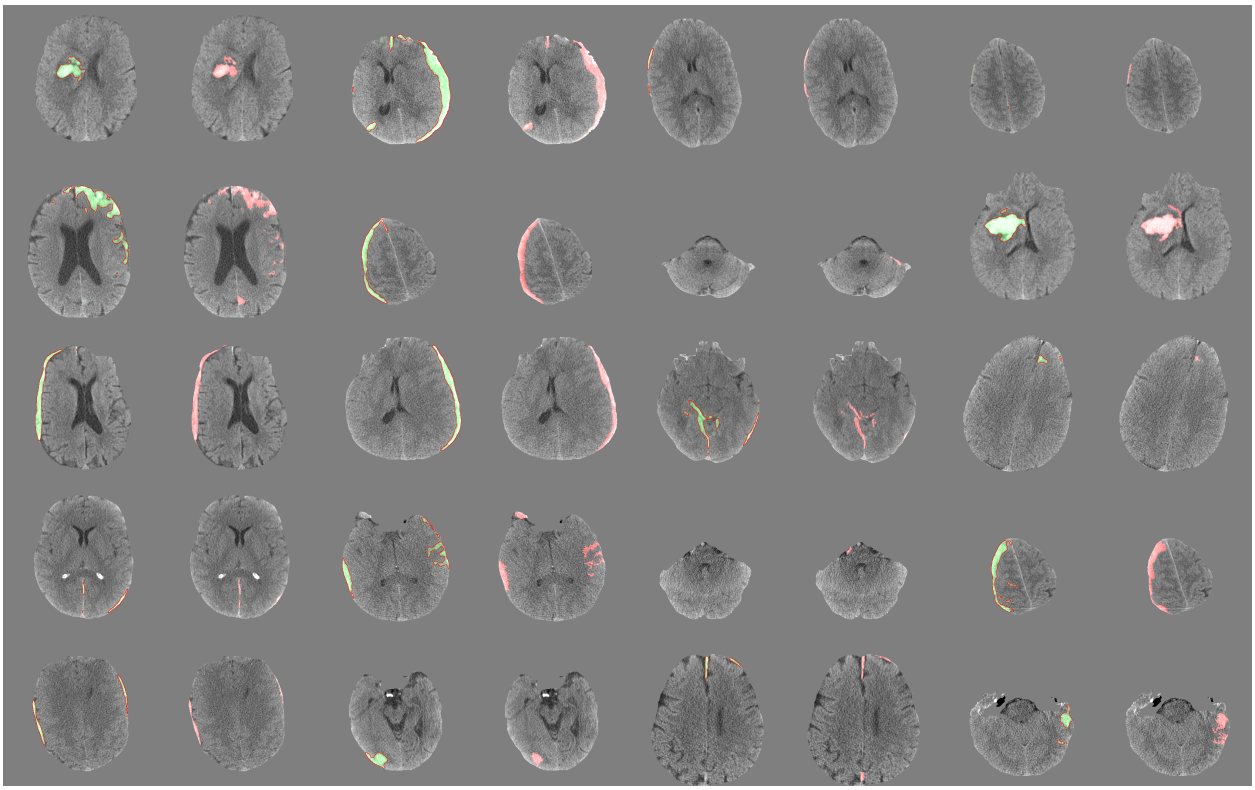

Figure 2.2: Visualization of PatchFCN segmentation. Each pair contains the PatchFCN output (left) and groundtruth labels (right). Results are randomly selected from the positive frames of the test set.

observe any gains using ImageNet. We re-weight the positive class loss by $\alpha = 3$ to balance the dominant negative class loss. The learning rate starts at 0.005 and decreases by a factor of 0.1 after 40% and 80% of the complete training iterations. At test time, we select $\beta = 3$ to ensure good overlap between adjacent sliding windows. To compute stack-level score, we select $p = 256$ in the L^p norm. All parameters were found by cross validation on the trainval set.

2.3 Experiments

Stack-level Benchmark with Human Experts

The first order task of hemorrhage detection is to determine whether a stack contains hemorrhage. We conduct internal as well as external validation for PatchFCN on stack-level as shown in Figure 2.3. The human expert is a neuroradiologist certified by the American Board of Radiology with 15 years of attending experience. The expert is instructed to examine each scan with the same level of care as a clinical scan. We allow the expert to take as much as

time as needed. The expert can modify their reads on scans before submitting final answers on the whole data set. The groundtruths are determined by at least one neuroradiologist with more than 10 years of neuroradiology attending experience.

Internal (Retrospective) Validation:

We report the ROC curve of PatchFCN on the test set and compare it with a human expert (15-year attending) in a retrospective setting where the test data was collected before the model development. Our single model AUC of 0.976 is competitive against the state-of-the-art 0.983 (single model) [4] and 0.993 (ensemble models) [18], while using much less training data. Our human expert has very low false positive rate 0.01 at 0.94 recall, better than the $(0.03, 0.90)$ of PatchFCN. Using both trainval and test data, our 4-fold cross validation AUC is 0.971 ± 0.006.

External (Prospective) Validation:

We collected a prospective test set of 200 scans after the model was developed. No further hyper-parameter adjustment was allowed in order to prevent overfitting to the test set. To minimize selection bias, we randomly select from all head CT scans performed from November to December 2018 using the Radiology Information System (RIS) SQL database in our hospital. The positive rate is 12.5%, which approximates the observed positive rates in emergency departments of many U.S. hospitals. For more details of data collection, please refer to the testing data section of chapter 3.2. Our ensemble model ($n = 3$) achieves an AUC of 0.966, which is competitive against the state-of-the-art 0.981 [4] and 0.961 [18]. PatchFCN approaches but does not exceed the human expert. Our best operating point is $(0.06, 0.92)$.

Pixel-level Evaluation

Apart from stack-level evaluation, we evaluate PatchFCN at pixel level because clinicians also want to know the location and volume of the bleeds for disease prognosis. Figure 2.2 visualizes the outputs of PatchFCN in comparison with the groundtruths. Results are shown on randomly selected positive frames in the retrospective test set.

On the retrospective test set, our model achieves pixelwise Dice score, Jaccard index, and average precision of 0.766, 0.620, and 0.785. In comparison, [4] reports Dice scores of 0.77 to 0.93 for a few types of hemorrhages they study. Our groundtruths are annotated pixelwise by senior neuroradiologists who specialize in TBI and include many subtle findings that could be easily missed by inexperienced radiologists. Using both trainval and test data, our 4-fold cross validation Dice score is 0.722 ± 0.027.

This section of the PatchFCN study is done after the study at Chapter 3. That's why the test data is identical.

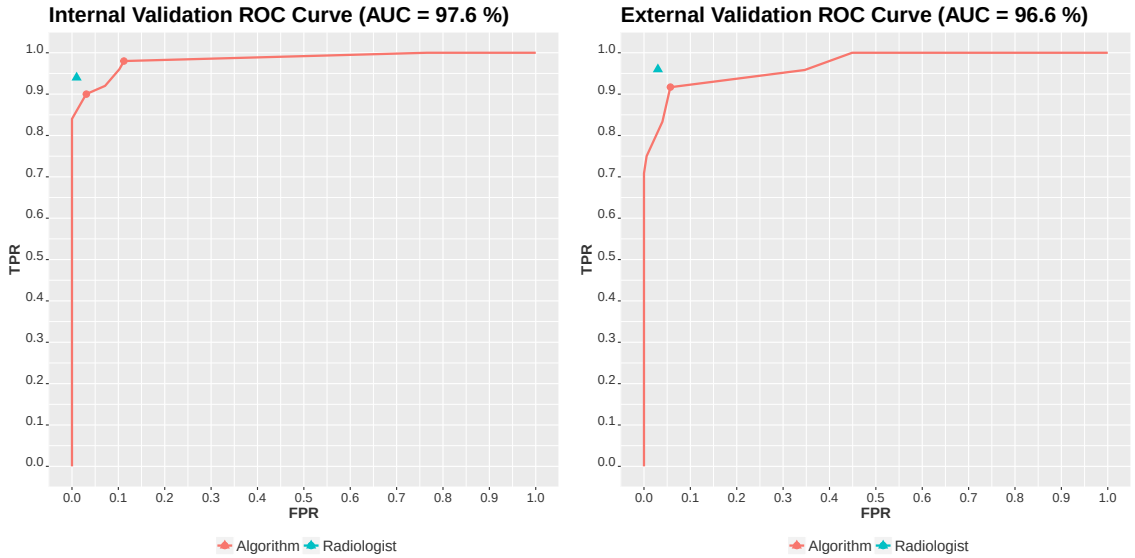

Figure 2.3: Internal and External Validation. We compare PatchFCN to an expert (neuroradiology attending with 15 years of experience) at stack level on retrospective and prospective test sets. PatchFCN achieves AUCs of 0.976 and 0.966 respectively, competitive with state-of-the-art systems that use much more labeled data. PatchFCN approaches but does not exceed the attending neuroradiologist.

Crop Size	80	120	160	240	480
Batch Size	144	64	36	16	4
Epoch	3600	1600	900	400	100
Dice	75.5	75.9	76.2	76.6	74.2
Jaccard	60.7	61.2	61.6	62.0	59.0
Pixel AP	78.5	78.1	78.5	78.5	75.9
Frame AP	87.8	89.3	89.8	89.9	87.8

Table 2.1: We benchmark PatchFCN on different patch sizes. Patch size 480 is the standard FCN that consumes whole images (baseline). As seen, PatchFCN consistently outperforms the baseline across a wide range of patch sizes on pixel and frame metric.

PatchFCN vs. FCN

Table 2.1 shows that PatchFCN consistently improves over standard FCN for pixel and frame by a healthy margin for a wide range of patch sizes. We report average precision (AP), Dice score and Jaccard index at pixel level with a threshold of 0.5. Note how PatchFCN is robust to patch size and maintains the performance even at a patch size of 80. We have tried even smaller sizes and observed a significant performance drop due to difficult optimization. To compare across different patch sizes, we choose the batch size to control the number of input

N	K	B	C	Epoch	Dice	Jaccard	PixelAP	FrameAP
16	1	16	240	400	76.6	62.0	78.5	89.8
8	2	16	240	200	76.4	61.8	78.5	89.7
4	4	16	240	100	74.7	59.6	77.3	87.7
2	8	16	240	50	57.5	40.3	67.6	81.4

Table 2.2: PatchFCN performance decreases with decreasing batch diveristy.

pixels per batch to be the same, and we choose the number of epochs such that the number of gradient steps are the same. We also ensure that all performances are saturated and training longer does not improve further.

What Makes PatchFCN effective?

Given the effectiveness of PatchFCN, we want to delve deeper to understand what makes patches so effective. We identify a few differences from standard FCN and study them by control experiments. For the following experiments, we define the batch size B, which is the product of N, the number of images per batch, and K, the number of patches per image. The batch size is defined this way because we sample patches from each of the image samples. PatchFCN has $K = 1$, $N = 16$, $B = 16$ and $C = 240$, where C is the crop size, whereas the standard FCN has $K = 1, N = 4, B = 4, C = 480$. We perform these analyses on the test split because it is larger and yields more stable performance. In this section, we control the number of input pixels and number of iterations the same way as in Section 2.3, unless otherwise stated.

Batch Diversity:

One possible advantage of PatchFCN is that we can fit a larger batch size and thus include more diverse data within any given GPU memory. To study the contribution of batch diversity, we control the batch size B and decrease the number of images N we sample patches from. Since $B = N \times K$, this means we sample more patches per image. As N decreases, we expect batch diversity to decrease as well. The default PatchFCN has $N = B$ and $K = 1$, which has the greatest diversity for any given B. By fixing the other hyperparameters, we can safely say the only difference here is the batch diversity. Note that we control the number of steps to be the same, so we decrease the number of epochs linearly with N.

Table 2.2 shows that decreased batch diversity results in lower pixel and frame-level performance. The breaking point is at $N = 2$, where the performance drops significantly from $N = 4$. We speculate that this is due to the use of batch normalization in residual networks[40]. This experiment demonstrates the importance of batch diversity for PatchFCN.

C	N	K	B	Epoch	Dice	Jaccard	PixelAP	FrameAP
64	16	1	16	400	66.4	49.7	65.8	74.5
120	16	1	16	400	72.5	56.9	74.7	82.2
240	16	1	16	400	76.6	62.0	78.5	89.9
360	16	1	16	400	73.9	58.6	73.4	85.8
480	16	1	16	400	74.1	58.8	75.6	87.7

Table 2.3: Context helps PatchFCN from $C = 64$ to 240, but not beyond.

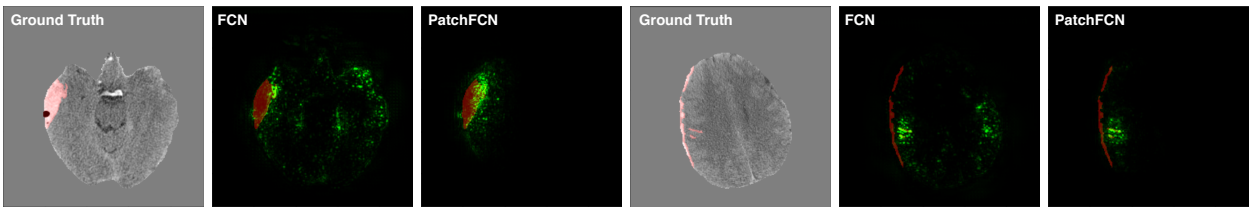

Figure 2.4: We visualize the gradients of PatchFCN with FCN in image space to see what cues the models rely on. Green speckles are the gradients and the burgundy regions are the selected ground truths for back-propagation.

How Much Context Does PatchFCN Need?

A trade-off of using patches is that we restrict the amount of context available to the network during training. Intuitively, one would think that more context is better. However, with limited amount of data, it is possible that less context could serve as an effective regularizer by forcing the prediction to rely on local information. To understand how much we lose/gain by having less context, we compare PatchFCN using different patch sizes while fixing the batch size and the number of steps (number of input pixels not the same here).

Table 2.3 shows that the improvement of context plateaus at patch size $C = 240$. Compared to $C = 64$, $C = 240$ is significantly better. However, increasing the patch size beyond 240 does not offer any more gain. We speculate that the improvement comes from the context regularization of patches, which helps in case of limited data. Overall, controlling context with patches is effective and allows the use of a larger and more diverse batch as in Table 2.2.

To qualitatively study what cues PatchFCN uses, we backpropagate the gradients from each hemorrhage region to the image space (see Fig.2.4). The gradient responses primarily come from the pixels not confidently predicted and correspond to the cues used for hemorrhage prediction. Fig. 2.4 shows that FCN captures long range dependencies that can easily overfit to limited data, while PatchFCN focuses on the local morphology and may generalize better.

C	B	Epoch	Dice	Jaccard	PixelAP	FrameAP
80	144	3600	69.4 (-6.1)	53.1 (-7.6)	74.9 (-3.6)	85.5 (-2.3)
120	64	1600	75.0 (-0.9)	60.0 (-1.2)	75.6 (-2.5)	88.7 (-0.6)
240	16	400	75.9 (-0.7)	61.2 (-0.8)	76.4 (-2.1)	89.8 (-0.1)

Table 2.4: Sliding window inference consistently outperforms fully convolutional inference (black numbers) for all patch sizes. The red numbers show the gap with sliding window inference.

Patch-based Sliding Window Inference:

At inference time, standard FCN applies on the whole image at once [20]. We hypothesize that this is sub-optimal for PatchFCN because the model is only trained on patches but has to take whole images at test time. That is why the default PatchFCN adopts sliding window inference to minimize the domain shift by letting PatchFCN evaluate patch by patch at test time. In Table 2.4, we show that sliding window inference consistently improves over fully convolutional inference for all patch sizes. Note that the gap is largest for the smallest crop size of 80, and decreases as patch size increases.

2.4 Summary

In this chapter, we propose PatchFCN – a simple yet effective framework for intracranial hemorrhage detection. PatchFCN approaches the performance of an expert neuroradiologist as well as performs competitively with the state-of-the-art at stack level. In addition, it localizes many subtypes of hemorrhages well and has strong pixel level performance. Analyses show that PatchFCN outperforms FCN by finding a good trade-off between batch diversity and the amount of context. We would demonstrate an improved version of PatchFCN in the next chapter.

Chapter 3

Expert-level PatchFCN

3.1 Introduction

In the previous chapter, we introduce PatchFCN and show validation on a relatively small training dataset. Since the scale of data is a major contributor to the success of deep learning, we ask the question: "how would the performance of PatchFCN improve if we scale up the training data?" To answer it, we collected a larger set of 4.4K exams from the hospitals affiliated with the University of California, San Francisco and conducted a thorough comparison with human experts.

With larger data, PatchFCN demonstrates the state-of-the-art accuracy among deep learning approaches, while concurrently localizing the abnormalities in the positive exams, including ones missed by experts. The performance of PatchFCN is above 2 out of 4 human experts, who are attending radiologists with 4 to 16 years of experience. As an exploratory study, we also demonstrate promising results for multiclass hemorrhage segmentation and detection.

3.2 Method

Model Architecture

We extend the PatchFCN presented in the previous chapter with a patch classification branch. The idea is to make the patch prediction completely learning-based and more robutst by detaching the patch prediction from the noisier pixel predictions. The entire system is shown in Figure 3.1.

This chapter is based on the PNAS submission titled "Expert-Level Detection of Acute Intracranial Hemorrhage on Head Computed Tomography using Deep Learning" done with Christian Hne, Pratik Mukherjee, Jitendra Malik, and Esther Yuh (submitted in May 2019). Statements about past work should be read with this context in mind.

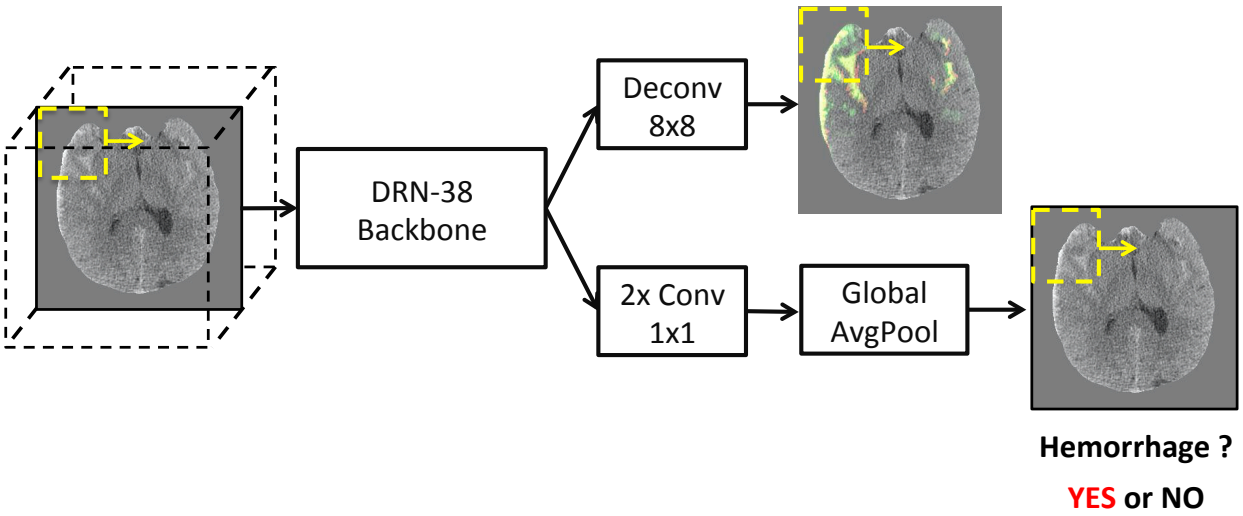

Figure 3.1: Expert-level PatchFCN system diagram. At the top pathway, we apply the PatchFCN as presented in chapter 2 including the inputs, backbone, and upsampling layer. At the bottom pathway, we apply two convolution followed by a global average pooling to obtain patchwise classification (bottom right image). The stack-level score is given by the maximum of patch-level scores within the stack. The green shows the prediction and red shows the ground truth annotation.

Data

Training Data

All patient data used in this study were collected retrospectively and de-identified, with no need for additional patient contact. Based on U.S. regulation 45 CFR 46.116(d) and the FDA, this study satisfied recommended conditions for ethically acceptable waiver of consent due to 1) minimal risk to patients, 2) no adverse effect on the welfare of patients, and 3) the impracticality of contacting very large numbers of subjects for a retrospective study. The study protocol was approved by the UCSF Committee on Human Research.

To develop the algorithm, we used a training set composed of 4,396 head CT scans performed at UCSF and affiliated hospitals (Table 3.1). This data set (UCSF-4.4K) consists of 1,131 exams positive for intracranial hemorrhage and 3,265 negative exams. The training dataset had a wide spectrum of sizes and types of hemorrhage as well as of imaging artifacts, and was collected from 4 different CT scanners from two major CT vendors (GE Healthcare and Siemens Healthineers) from 2010-2017. Each exam consisted of a 3D stack of 27-38 transverse 2D images through the head acquired on 64-detector-row CT scanners. Pixelwise labels for acute intracranial hemorrhage were verified by two ABR certified radiologists with CAQ in Neuroradiology.

https://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM566948.pdf

Split	CT Manufacturer	Number of Exams	Hemorrhage	Number of Exams
Training	GE	1589	Positive	1131
Training	Siemens	2807	Negative	3265
Testing	GE	14	Positive	25
Testing	Siemens	186	Negative	175

Table 3.1: Training and test data by CT manufacturers and by positive/negative count.

Testing Data

To validate the algorithm, we collected a separate test set of 200 head CT scans performed at the same hospitals in November-December 2017 (Table 3.1). In formulating the test set, we aimed for an overall 10 to 15% positive rate for acute intracranial hemorrhage that approaches the positive head CT rate in many busy acute-care hospitals. We also wished to evaluate the algorithm on the initial head CT exam only, and to exclude follow-up head CT exams performed during the same hospitalization following neurosurgical interventions such as hemicraniectomy or craniotomy. We also aimed to include within the test set a substantial number of positive exams that would include a diverse spectrum of possible intracranial hemorrhage patterns, while maintaining an overall low positive head CT rate that would simulate observed rates in current clinical practice. We needed to control the overall test set size, such that each adjudicating radiologist could interpret the entire set of 200 head CT exams within a total of 5 days when working at an average clinical pace. Finally, we wished to minimize selection bias in the process of selecting cases for the test set. To accomplish these goals for the test set, we used the following approach. The exams were identified from the Radiology Information System (RIS) Structured Query Language (SQL) database. Using the RIS, we randomly selected 150 head CT exams ordered from November to December 2017 that excluded reference to a prior craniectomy or craniotomy; and for which no prior or follow-up head CT exam was found for that patient during the same hospitalization. We also randomly selected 50 head CT exams with no reference to prior craniectomy or craniotomy, and no prior head CT exam during the same hospitalization, but with at least one follow-up head CT scan performed during the same hospitalization. Since most CT scans with no follow-up CT scan during the same hospitalization are negative for an acute intracranial abnormality, while many (but not all) CT scans with at least one follow-up CT scan performed during the same hospitalization contain a significant acute intracranial finding, we estimated that this strategy would yield an overall 10 to 15% proportion of positive head CT exams for acute intracranial hemorrhage, while avoiding the need to view the actual images. Using this approach, the actual test set of 200 exams contained 25 positive and 175 negative exams for acute intracranial hemorrhage, for an overall 12.5% positive rate that approximates the observed positive head CT rate in many hospitals. The skull stripping algorithm failed on one head CT exam, which was replaced by another exam from the same time period using the same approach. The test set did contain a larger proportion of Siemens

Class	0	1	2	3	4
Hemorrhage	None	SDH	EDH	Contusion, ICH, TAI	SAH, IVH
Pixel Ratio	0.996	3.5×10^{-3}	3.2×10^{-4}	2.2×10^{-5}	7.1×10^{-4}
Exam Ratio	0.686	0.196	0.026	0.152	0.232

Table 3.2: Multiclass Exploratory Data.

CT exams compared to the CT vendor distribution in the UCSF-4.4K training data set, owing to the larger number of head CT exams performed on Siemens CT scanners as part of the acute head CT workflow in place at Zuckerberg San Francisco General Hospital and Trauma Center (ZSFG) during the November-December 2017 time period.

Multiclass Training Data

To explore the potential of PatchFCN in multiclass setting, we collected an expanded set of multiclass hemorrhage data that comprises of 4766 scans from GE and Siemens scanners. The exams are conducted and labeled following the same protocol as described earlier. We label each pixel with its hemorrhage type label. We define the hemorrhage classes by clinical criteria shown in Table 3.2. The pixel and exam ratios of each class indicate the proportion of positive pixels/exams with the class of hemorrhage present. Note that the positive-class pixels are extremely rare compared to the negatives. The scarcity of foreground pixels in conjunction with low-contrast noisy images makes both pixel and exam-level prediction challenging.

Data Preprocessing

The skull and face were removed from CT images using a series of image processing techniques, including thresholding to identify skull and facial bones, followed by a series of close, open and fill operations to retain only the intracranial structures. This enhanced privacy of the data, as individuals could in theory be identified through surface rendering of facial soft tissue pixels present in the original data. It also makes the problem easier for the network as it only needs to model the intracranial structures.

Data Availability

The data used to train and test the machine learning models are administered by the University of California (California Code Regs. title. 22 Section 70751). The data set in in its entirety is not currently publicly available, but a subset may be available for research, subject to approval of the UCSF Committee on Human Research.

Multiclass Architecture

We conducted an exploratory study on the multiclass prediction of hemorrhage types at the pixel and exam levels. The model output layers are re-designed for the tasks as follows: 1) the pixel classifier has N + 1, instead of 2, output channels, where N is the number of hemorrhage classes. 2) the stack classification branch has 2(N + 1) outputs for the N hemorrhage classes and the combined positive class. This design is motivated by the observation that the classes are mutually exclusive at the pixel level (i.e., each pixel is a member of only one class, or subtype, of hemorrhage) but not at the exam level (i.e., each exam can contain multiple classes of hemorrhage).

Implementation Details

The network backbone architecture was Dilated ResNet 38 [40], and all hyperparameters were developed on the UCSF-4.4K training set described below. We optimized cross entropy loss with stochastic gradient descent (SGD) and a momentum of 0.99. The learning rate was decreased by 0.1 every 160 epochs. To control class imbalance, we sampled 30% of the patches from positive images in each training mini-batch and up-weighted the positive pixel loss by a factor of 3. At training time, the backbone and the pixel prediction branch (one up-convolution layer) were trained at an initial learning rate of 10-3 for 400 epochs. Both of these were then fixed, and the patch classification branch (conv + batchnorm + ReLu + conv layers) was trained for 40 epochs. Finally the entire model was jointly fine-tuned for 30 epochs at a learning rate of 5 x 10-5. At inference time, adjacent patches were sampled at 2/3 overlap with each other. The pixel predictions in each patch were mapped to image space and averaged to yield the final prediction. The stack classification score was taken as the maximum patch classification score in the stack. The model evaluates each stack within one second on average.

3.3 Evaluation and Benchmark with Radiologists

Evaluation Protocol

To evaluate model performance, the deep learning algorithm was executed exactly once on the test set of 200 CT exams, with no adjustment of hyperparameters that had been selected during the algorithm development phase. This excluded the possibility of any overfitting to the test data, so that the reported performance should match the models true performance very well. For each scan in the test dataset consisting of 200 CT exams, the algorithm indicates both pixel-level and exam-level probabilities (continuous from 0 to 1) for the presence of intracranial hemorrhage. Although some patients underwent two or more head CT exams during the same hospitalization, it was ensured that each patient appeared at most once in either the training set or the test set, but not in both. We calculated the ROC for the deep learning algorithm to identify the presence of acute intracranial hemorrhage

on each CT exam, compared to the gold standard. The gold standard for interpretation of CT scans in the test set as positive or negative for acute intracranial hemorrhage consisted of a careful consensus interpretation by two ABR certified neuroradiologists with CAQ in Neuroradiology, one with 15 years and the other with 10 years of attending-level experience in interpretation of head CT exams.

Four ABR-certified practicing radiologists each reviewed the 200 CT exams in the test set. One radiologist had 2 years of subspecialty fellowship training and a CAQ in Neuroradiology, with 15 years of attending neuroradiologist experience. The others had 4, 10, and 16 years of experience in private and/or academic general radiology practice, including interpretation of head CT. Radiologists were asked to indicate whether each scan was more likely positive or more likely negative for acute intracranial hemorrhage, a binary decision, in contrast to the continuous probability for hemorrhage provided for each exam by the algorithm. Radiologists time to evaluate each scan was not limited. Radiologists were instructed to interpret all CT scans carefully, using conventions, such as the duration of time spent on each scan, and level of care in interpreting each scan, that would be consistent with U.S. standard-of-care clinical practice. Radiologists were able to return to prior CT scans and to modify their interpretations of exams they had seen earlier in the data set. Radiologists were not aware of the overall ratio of positive to negative CT exams. We calculated the sensitivity and specificity of each radiologist to detect whether or not there was acute intracranial hemorrhage on each CT exam, compared to the gold standard.

Benchmark and Visualization

Hemorrhage Detection

Figure 3.2 shows that our system PatchFCN performance exceeded that of 2 of 4 ABR-certified radiologists, with a receiver operating characteristic (ROC) with area under the curve (AUC) of 0.991 0.006 for identification of acute intracranial hemorrhage, referenced to the gold standard consensus interpretation of two ABR-certified neuroradiologists with Certificate of Added Qualification (CAQ) in Neuroradiology. In addition, PatchFCN achieved 100% sensitivity at specificity levels approaching 90%, making this a suitable screening tool with an acceptably low proportion of false positives.

Hemorrhage Segmentation

Figures 3.3A-L show examples of PatchFCN localization of acute intracranial hemorrhage in acute aneurysm rupture, hemorrhagic stroke, subacute traumatic brain injury, and acute traumatic brain injury. Of note, Figures 3.3J-L show an isodense subdural hemorrhage, and demonstrates that PatchFCN algorithm cannot rely solely on hyperdensity relative to brain in order to identify acute hemorrhage, but must also use other more subtle features, as do experienced radiologists. Figures 3.4A-O demonstrate all positive cases in the 200-exam test set that were missed by at least 2 of 4 radiologists.

Class	1	2	3	4	Combined
Types	SDH	EDH	Contusion/ICH/TAI	SAH/IVH	All
ROC Area %	95.4 ± 1.0	94.0 ± 1.6	93.4 ± 0.7	95.6 ± 0.6	98.2 ± 0.4

Table 3.3: Exam-level Multiclass Hemorrhage Detection. SDH - subdural hematoma. EDH - epidural hematoma. ICH - intracerebral hematoma. TAI - traumatic axonal injury. SAH - subarachnoid hemorrhage. IVH - intraventricular hemorrhage.

Cross Validation

To confirm reproducibility of results, we conducted 4-fold cross-validation experiments. We randomly split the UCSF-4.4K training data into 4 subsets. For each of 4 experiments, 3/4 of the UCSF-4.4K set was used for training and 1/4 was held out as a test set. The 4 resulting ROC curves demonstrated AUC values of 0.978 *pm* 0.003, which were slightly lower than the AUC of 0.991 based on training on the full UCSF-4.4K set. However, the small standard deviation of 0.003 demonstrates reproducibility of results. Regarding localization accuracy, the algorithm achieved an average Dice coefficient of 0.75 on the 4-fold cross-validation experiments.

Multiclass Exploratory Study

Table 3.3 shows that PatchFCN achieves competitive exam-level multiclass detection on our expanded exploratory dataset, while maintaining the strong two-class results on 4-fold cross validation. The results are reported as mean one standard deviation. We note that the exam-level prediction of each class (including the combined class) is made with an independent binary classifier at the output layer. Figure 3.5 shows examples of multiclass segmentation by the algorithm and by a neuroradiologist.

Discussion

We report a deep learning algorithm with accuracy comparable to that of radiologists for the evaluation of acute intracranial hemorrhage on head CT. We show that deep learning can accurately identify diverse and very subtle cases of a major class of pathology on this workhorse medical imaging modality. Head CT interpretation is regarded as a core skill in radiology training problems, and the performance bar for this application is accordingly high, with the most experienced readers demonstrating sensitivity/specificity between 0.95 and 1.00.

In this study, we demonstrate, to our knowledge, the highest accuracy levels to date for this application by using a PatchFCN with strong supervision and a relatively small training data set, compared to prior work relying on weaker supervision using exam- or image-level labels [36, 18, 25, 5] or Mask R-CNN [4]. We show that FCN with pixel-level supervision is well-suited to this application, in which poorly-marginated abnormalities of

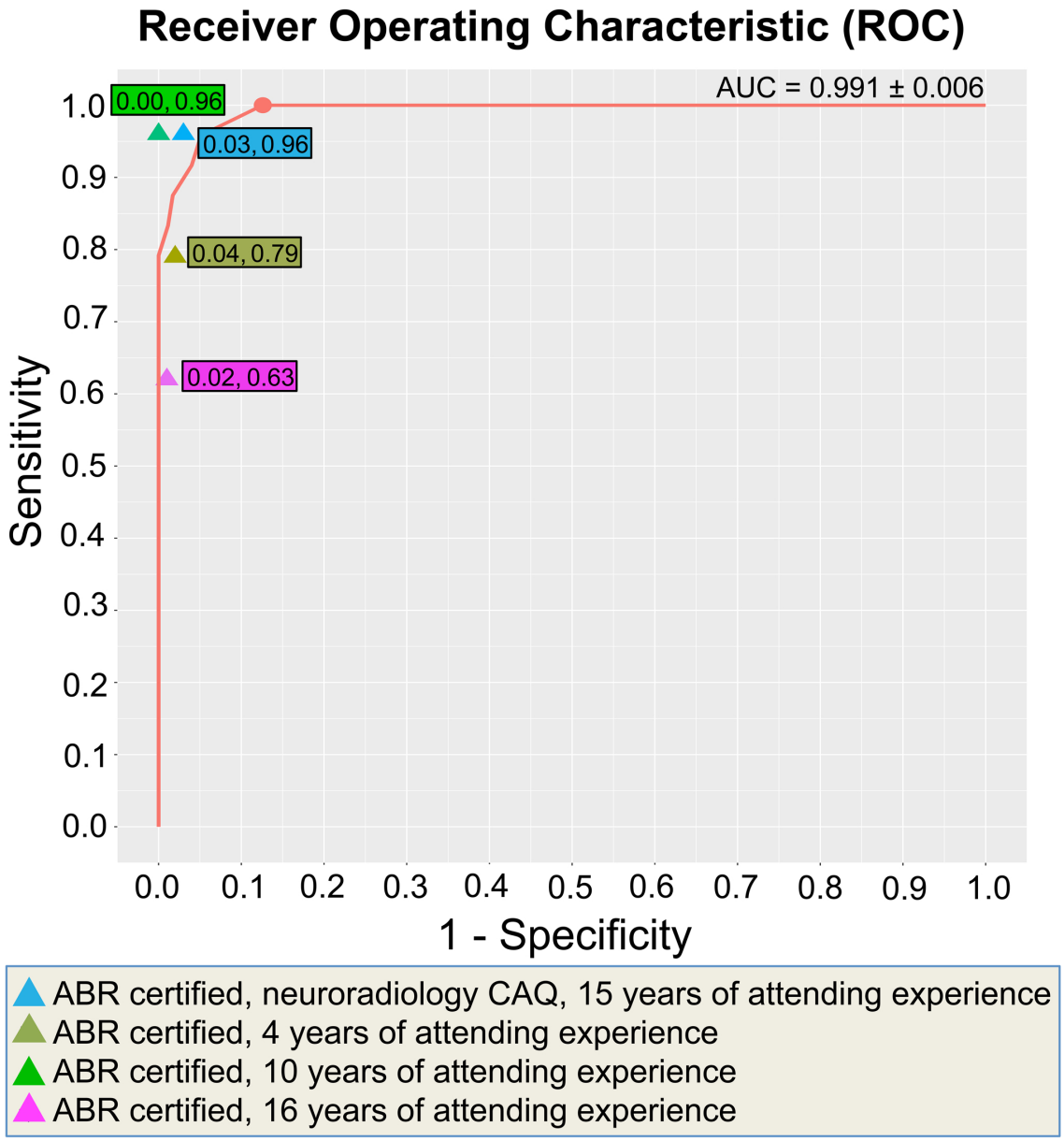

Figure 3.2: Receiver operating characteristic (ROC) for the deep learning model to predict the presence of acute intracranial hemorrhage on 200 head CT exams. The algorithm achieved an area under the curve (AUC) of 0.991 0.006 referenced to the gold standard (consensus interpretation of two ABR-certified neuroradiologists with Certificate of Added Qualification (CAQ) in Neuroradiology). Algorithm performance exceeded that of 2 of 4 American Board of Radiology (ABR) certified radiologists with attending-level experience ranging from 4 to 16 years. In addition, PatchFCN achieved 100% sensitivity at specificity levels approaching 90%, making this a suitable screening tool for radiologists based on an acceptably low proportion of false positives.

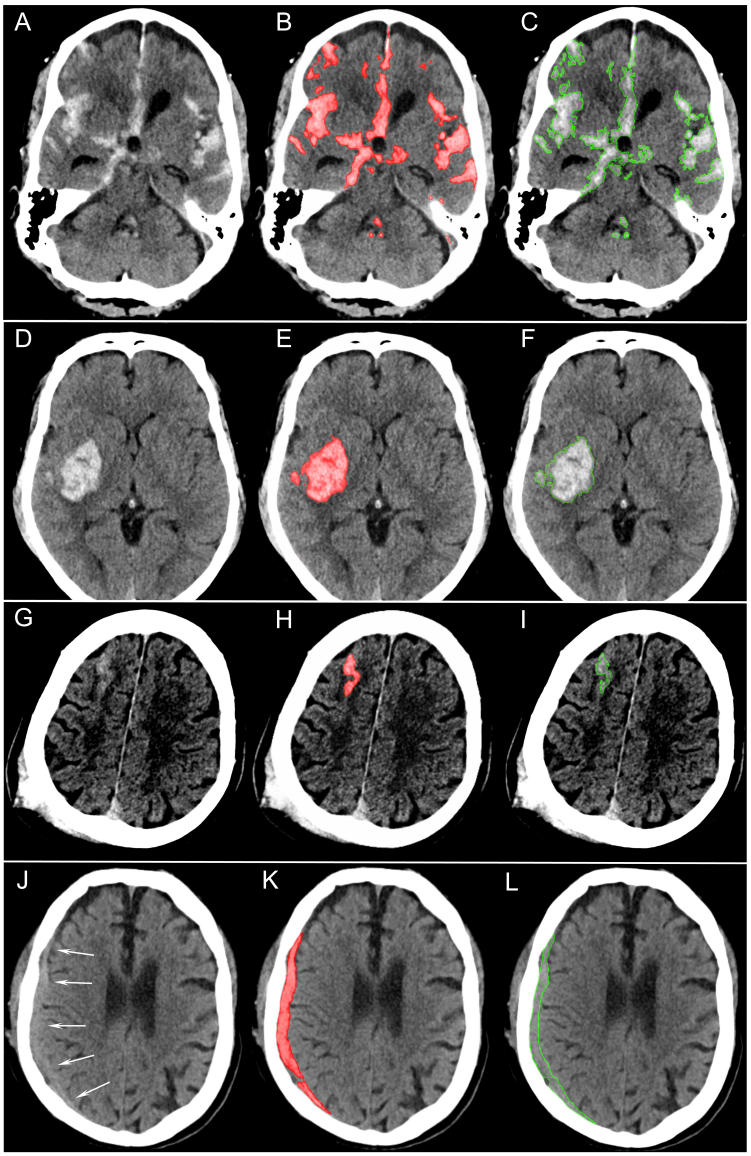

Figure 3.3: Patch-based Fully convolutional neural network (PatchFCN) segmentation of acute intracranial hemorrhage. A-C, Subarachnoid hemorrhage (SAH) due to aneurysm rupture. D-F, Acute intracerebral hemorrhage. G-I, Traumatic SAH (missed by one of 4 radiologists) and J-L, isodense subdural hematoma (SDH). J-L, may represent either an acute SDH in the setting of coagulopathy, or a subacute SDH at 2 to several days after injury. Because isodense subdural hematomas are not brighter than the adjacent brain parenchyma, radiologists identify these by recognizing the absence of sulci and gyri within the isodense collection. In J-L, the SDH is detected despite its isodensity to gray matter, showing that the deep learning algorithm does not rely solely on hyperdensity, but also uses other features to identify hemorrhage. A,D,G,J, Original images. B,E,H,K, Original images with red shading of pixel-level probabilities over 0.5 (on a scale of 0 to 1) for hemorrhage, as determined by the PatchFCN; pixels with probability below 0.5 were unaltered from the original images. C,F,I,L, Neuroradiologists segmentation of hemorrhage using green outline.

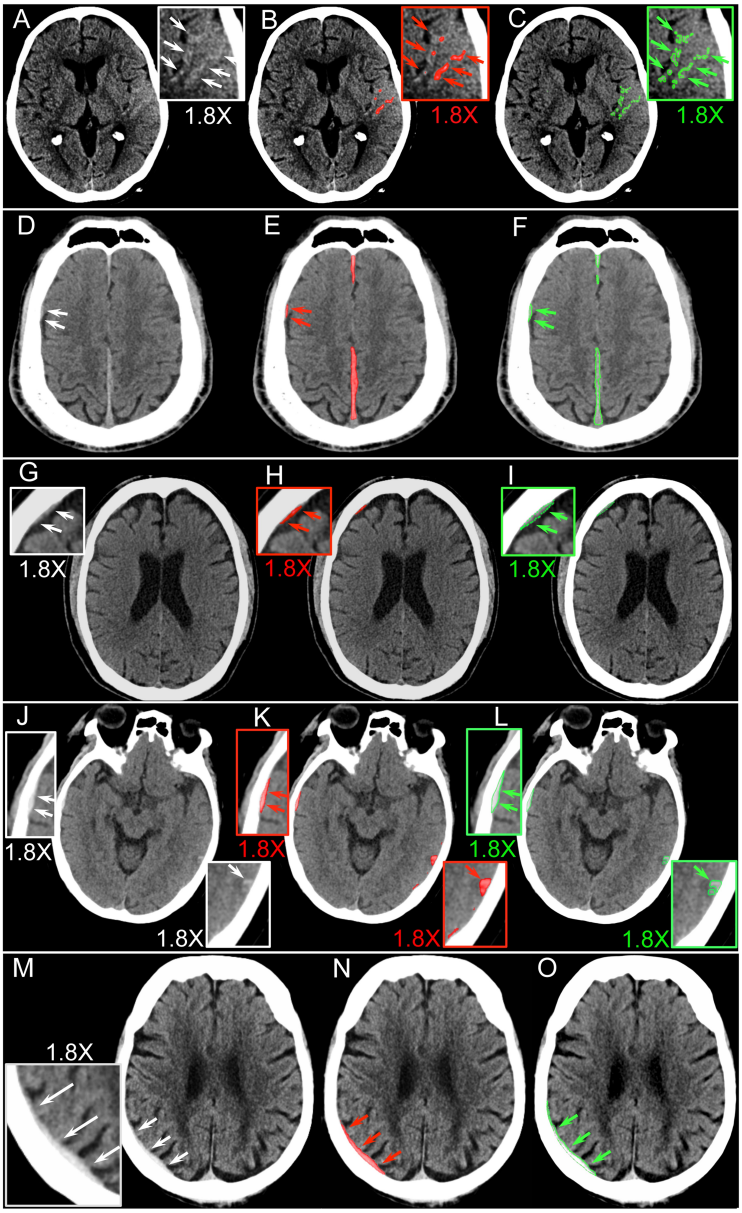

Figure 3.4: Five cases judged negative by at least 2 of 4 radiologists, but positive for acute hemorrhage by both the algorithm and the gold standard. A-C, Small left temporal subarachnoid hemorrhage (SAH), D-F, small right posterior frontal and parafalcine subdural hematomas (SDH), G-I, small right frontal SDH, and J-L, small right temporal epidural hematoma and left posterior temporal contusion were each called negative by 2 of 4 radiologists. M-O, was called negative by all 4 radiologists but contained a right parietal SDH identified by both the algorithm and by the gold standard. A,D,G,J,M, Original images. B,E,H,K,N, Algorithmic delineation of hemorrhage with pixel-level probabilities over 0.5 colored in red. C,F,I,L,O, Neuroradiologist segmentation of hemorrhage using a green outline.

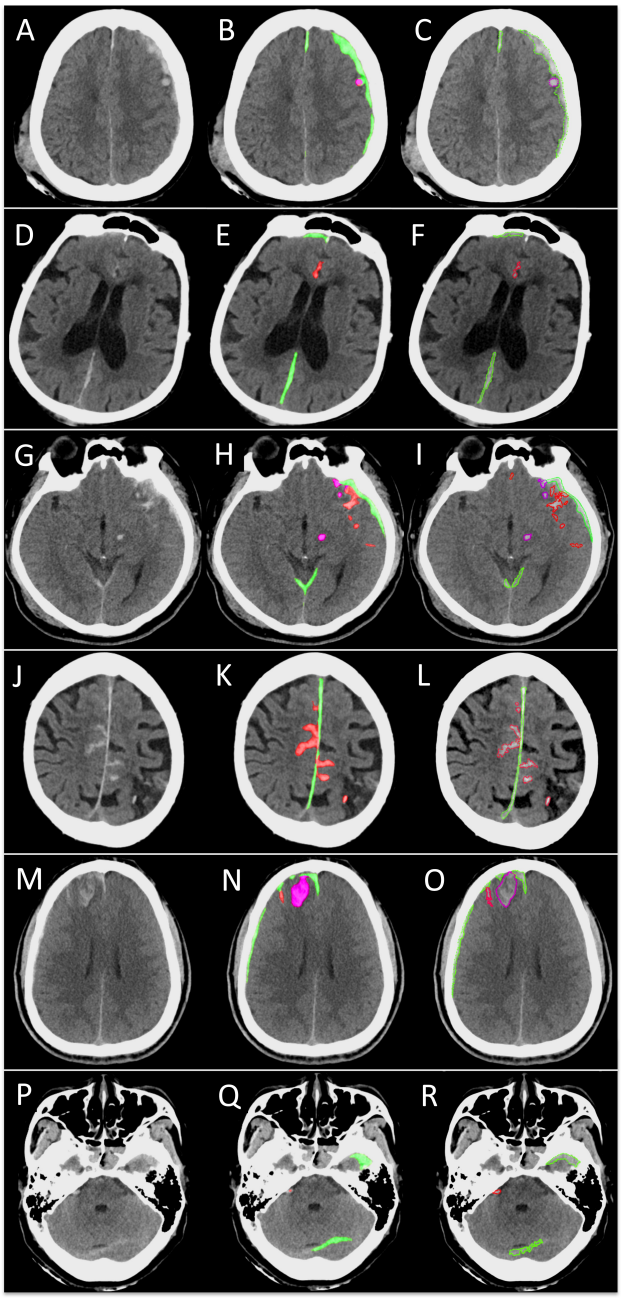

Figure 3.5: Examples of multiclass segmentation by the algorithm and by an expert. A-C, Left holohemispheric subdural hematoma (SDH, green) and adjacent contusion (purple). D-F Right frontal and posterior parafalcine SDHs (green) and anterior interhemispheric subarachnoid hemorrhage (SAH, red). G-I, Tentorial and left frontotemporal SDH (green) and subjacent contusion (purple) and SAH (red), in addition to shear injury in the left cerebral peduncle (purple). J-L, Parafalcine SDH (green) with surrounding SAH (red). M-O, Several right frontal SDHs (green) with subjacent contusion (purple) and SAH (red). P-R, Small left tentorial and left anterior temporal SDHs (green) and right cerebellopontine angle SAH (red). A,D,G,J,M,P, Original images. B,E,H,K,N,Q, Algorithmic delineation of hemorrhage with pixel-level probabilities over 0.5 colored in red (SAH), green (SDH), and contusion/shear injury (purple). C,F,I,L,O,R, Neuroradiologist segmentation of hemorrhage.

widely varying sizes and morphologies, such as hemorrhage, need to be both detected and localized. Improving on previous reports [36, 18, 25, 5, 4], we achieve 100% sensitivity for acute hemorrhage detection at 90% specificity, which represents an acceptable rate of false positives for clinical screening purposes.

In addition, motivated by the clinical need to identify and localize, in most cases, a very sparse foreground (e.g., examples of hemorrhage in Figure 3) with high sensitivity, we applied the PatchFCN from chapter 2 that was informed by just the right amount of local information [16]. Specifically, limitation of the network evaluation of each 2D image on any single pass to a subset or patch of the 2D image for modeling x-y-axes context consistently outperformed evaluation of the entire 2D image on pixel and exam-level. We surmise that a reason for this may be that deeper models with a massive number of free parameters may overfit to less relevant distant information in large input images in the setting of a limited data set size. Similarly, we use 3 consecutive frames as inputs to the network following PatchFCN described in chapter 2. This is based on the finding that a network informed by 3 consecutive transverse (i.e., axial) images (image under evaluation, and flanking images immediately superior and inferior) was as accurate for pixel and exam-level classification as a network that employed 5 or more consecutive images, sparing the need for learning even more context with 3D-FCN and avoiding the problem of overfitting to too large a context. 3D-FCN takes in the entire 3D volume, and was demonstrated to achieve accuracy levels exceeding that of human experts for classification of OCT exams [6]. For the current application, in which a single small localized area of less than 100 pixels on a single image may represent the sole abnormality in a 3D volumetric stack comprising approximately 30 images and a million pixels, we found in chapter 2 that the theoretical advantage of taking in more global context was outweighed by the advantages of 1) forcing the network to consider an intermediate amount of spatial context, both in-plane and in the craniocaudal direction, and 2) larger batch diversity to stabilize training through the use of batch normalization in deep networks.

The detection of these tiny acute hemorrhages can be of life-saving importance, since an SAH of less than 100 pixels may be the only evidence of a sentinel bleed from a cerebral aneurysm. If the abnormality is missed and the patient sent home from the Emergency Department without treatment of the underlying aneurysm, he or she is at risk of death or long-term disability when the aneurysm ruptures. Indeed, in half of cases when the emergency CT scan is interpreted as negative but the patient is later found to have a cerebral aneurysm, the acute hemorrhage is found in retrospect to have been definitely or probably present on the head CT but missed by the radiologist [23]. Similarly, a tiny EDH that is missed on an emergency CT scan after head trauma has the potential to rapidly expand and kill the patient within hours in the absence of neurosurgical evacuation of the hematoma. While 100% accuracy for acute hemorrhage detection is desirable under these circumstances, unfortunately, the human experts who provide the training for deep learning algorithms are fallible and there is no perfect gold standard for intracranial hemorrhage detection currently available to better train these algorithms. However, by learning from the inputs of multiple human experts, an accuracy level exceeding any single human expert may become feasible.

Our exploratory multiclass results demonstrate higher levels of classification accuracy (93.4% to 95.6%) across the entire spectrum of hemorrhage types than has previously been achieved [5], and with no loss of overall hemorrhage detection accuracy, despite being performed at pixelwise resolution. These multiclass results also constitute the first prospective demonstration of accurately classifying and segmenting EDHs, which can be difficult to distinguish from SDHs since both represent extra-axial hematomas. This is a crucial distinction, given the clinical importance of accurate early diagnosis of EDH, as discussed above.

To address the need for both accurate exam-level classification and concurrent localization of abnormalities at the pixel level, we used a single-stage network for joint segmentation and exam-level classification, which enjoys the advantages of 1) only one network for both segmentation and exam classification instead of two at both training and test time, and 2) significant feature sharing between segmentation and classification networks. In general, it is beneficial to share the representation between correlated tasks, which saves computation and also serves as an effective regularization method [12]. Figure 3.1 summarizes our hemorrhage detection system architecture.

In summary, we demonstrate a deep learning algorithm for detection and localization of acute intracranial hemorrhage on head CT, based on a strong supervision approach and a relatively small training data set. We show performance that is comparable to highly-trained experts. Beyond the key clinical tasks of classification of head CT exams as positive or negative for abnormalities, PatchFCN will be useful for deriving quantitative biomarkers from CT and other radiological exams. Rudimentary size measurements for intracranial hemorrhage already play a role in practice guidelines for the management of acute hemorrhagic stroke (ABC/2 method for quantifying intracerebral hematoma [14, 35]), acute aneurysmal subarachnoid hemorrhage (Fisher Grade [10]), and acute TBI (Marshall [24] and Rotterdam scores [21] and criteria for performing decompressive hemicraniectomy [3].) Even these crude measurements are subjective and can be time-consuming to obtain [28]. Improved quantitative information has not been explored due to the impracticality of obtaining these for large datasets, particularly for poorly-marginated ill-defined abnormalities such as subarachnoid and multifocal intracranial hemorrhage, both of which are common in clinical practice. The ability to identify, localize, and quantify features is likely to provide more granular data for research into therapies, prognosis, risk stratification, best treatment practices, and the cost effectiveness of imaging tests.

Chapter 4

Cost-Sensitive Active Learning

4.1 Introduction

Clinical applications set very high bars for machine learning algorithms, because any misdiagnosis could impact treatment plans and gravely harm the patient. For example, general radiologists are known to read the traumatic brain injury (TBI) on head CT scans at a misinterpretation rate of 2.7%-5% [17]. In chapter 3, we collect a densely labeled dataset of 4.4K scans and demonstrate an expert-level PatchFCN for this task. To go further and outperform human experts, it is impractical to continue labeling new exams naively, because most of the exams would already be correctly predicted by the model. What we want to label are those exams the model is not confident about.

Active learning (AL) aims to address the paucity of labeled data by reasoned choice of which available unlabeled examples to annotate [31, 39, 32, 19, 22]. The use case is to grow the annotated data without active learning until the performance starts to saturate. Then active learning comes in to select the most informative examples from a large unlabeled pool to continue the performance gain.

A limitation of many prior studies of AL is that they validated AL only in a core-set selection setting, [29] rather than demonstrating its utility in growing the labeled data, and also did not attempt to model the cost of labeling [31, 39, 22]. However, the potential value/use of AL is not in achieving comparable performance with less data, but in improving the model while also minimizing labeling costs. On other problems it has been shown that labeling costs vary greatly from one example to another [31, 30, 37]. In the case of intracranial hemorrhage, we observe that times needed for pixelwise labeling vary up to 3 orders of magnitude for different cases (See Fig. 4.2) much more than what is typically seen in computer vision or natural language processing. We believe similar phenomena may well exist in other medical imaging domains. Most AL studies to date select examples without addressing this wide variation in labeling time [39, 32, 19, 29, 22].

This chapter is based on the work [15] work done with Christian Hne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Statements about past work should be read with this context in mind.

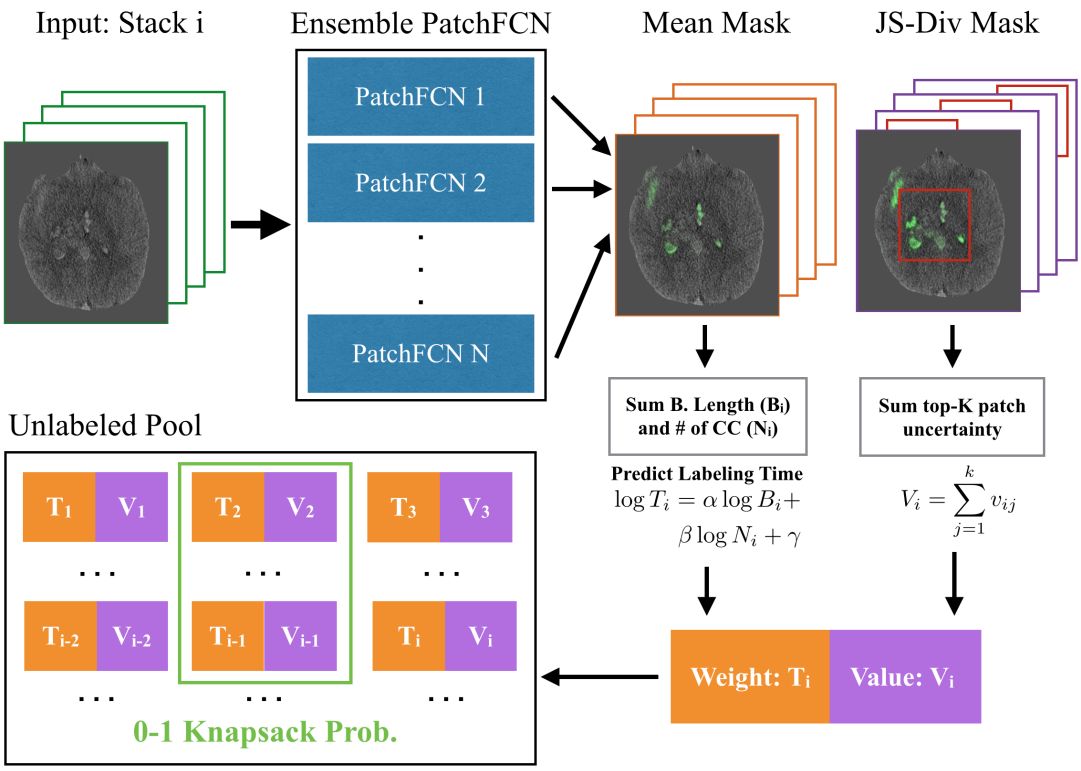

Figure 4.1: Overview. First, the stack runs through the ensemble PatchFCNs trained on the seed set S, which produces the mean hemorrhage heatmap and the Jensen-Shannon (JS) divergence uncertainty heatmap. From the mean hemorrhage heatmap, we apply multiple thresholds to compute the mean boundary length B_i and number of connected components N_i. Our log-regression model then takes B_i and N_i to predict the stack labeling time T_i. The sum of uncertainty of the top-K uncertain patches is defined to be the stack uncertainty V_i. Given any fixed labeling budget(time) Q, we treat each stack in the unlabeled pool as an item of weight T_i and value V_i. The optimal set of items for annotation is obtained by solving a 0-1 Knapsack problem with dynamic programming.

In this paper, we propose a cost-sensitive AL system by combining the query-by-committee approach with labeling time prediction for each example [32]. Our uniform-cost AL system compares favorably with the state of the art [39], while the cost-sensitive system gives a further boost under labeling time constraints. All experiments are conducted on a dataset of 1247 exams (29095 frames), which is about two orders of magnitude larger than standard MICCAI segmentation datasets [33, 43]. Moreover, our system is simpler, faster, and uses less memory than earlier works [39, 29]. Through the example of intracranial hemorrhage detection, we demonstrate the potential of cost-sensitive active learning to scale up medical datasets efficiently.

4.2 Cost-sensitive Active Learning

Let us define our active learning problem as follows: given a labeled seed set S and an unlabeled pool set U, find a small subset P from U for labeling that maximizes a suitable test set metric. Our system which is depicted in Fig. 4.1 estimates an uncertainty score for each example (see Sec. 4.2) and the labeling time (see Sec. 4.2). The goal is to select the set of examples such that the sum of their uncertainty is maximized under the constraint that the total estimated labeling time stays within a given budget. The optimal selection of items reduces to the well-known 0-1 Knapsack problem, which can be solved with dynamic programming.

Uncertainty Measure

Uncertainty (or informativeness) is at the core of active learning techniques. It can be estimated by single model outputs [19] or a committee of models [32]. The idea of query-by-committee (QBC) is to run multiple models on the same example and use their disagreement to estimate uncertainty. Experimentally, we found that QBC consistently works better than single-model uncertainty. Within the QBC framework, we have tried various uncertainty measures and found the Jensen-Shannon (JS) divergence to work best. Concretely, let's assume we have N models in the committee and the output distribution of model i is P_i. The JS divergence is then defined as:

$$JS(P_1, P_2, ..., P_N) = H(\frac{1}{N}\sum_{i}^{N} P_i) - \frac{1}{N}\sum_{i}^{N} H(P_i) \tag{4.1}$$

where H is the entropy function.

We average all pixelwise uncertainties within each patch to obtain the uncertainty of a patch. The stack uncertainty is obtained by averaging the top K uncertain patches within the stack. The choice of K is a balance between taking the max ($K = 1$) or the mean ($K = \infty$) of the whole stack. In all AL experiments in this paper, we set $K = 200$ and number of models $N = 4$. We have tried larger N but didn't gain any performance. Visualization of such uncertainty can be found in Fig. 4.5.

Labeling Time Prediction

First, we need to ask what is the optimal unit of labeling – patch, frame or stack? Employing our neuro-radiology expertise, we settled on labeling stacks. While labeling patches/frames may seem more effective from a machine learning perspective, it comes with a severe overhead, i.e. the whole stacks need to be retrieved and examined by radiologists anyway. Therefore, it is less efficient than labeling the stacks.

To apply active learning in practice, we need to ensure it actually saves labeling cost or efforts. This is crucial as per-stack labeling times in our data span 3 orders of magnitude. We

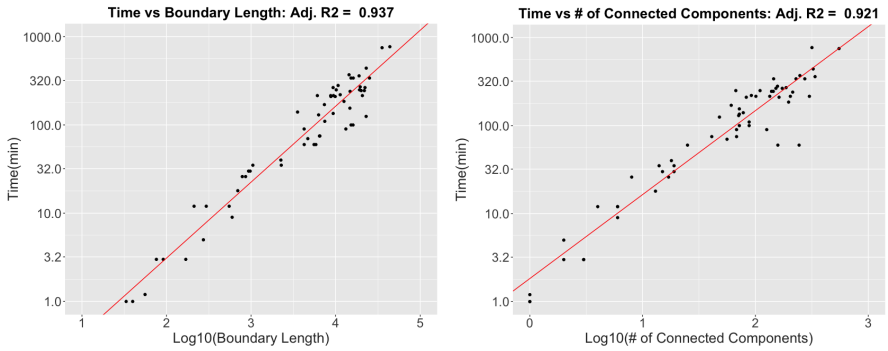

Figure 4.2: Left: Time vs Log(Boundary Length). Right: Time vs Log(Number of Connected Components). Both plots show the goodness of our linear fit and the normality of residuals after the log transform. Note that the y-axis is actually displayed in log-scale.

utilize linear regression to predict the log labeling time $\log t$ based on two features: 1) mask boundary length B, and 2) number of connected components M under the log-transform.

$$\log t = \alpha \log B + \beta \log M + \gamma \tag{4.2}$$

Fig. 4.2 shows the effectiveness of our log-transform and the goodness of fit on both features. 61 data points were used to fit the linear model, which we found to be sufficient. In order to compute the features at test time we use the pixelwise predictions of our network. We also tried using deep FCN features from an intermediate layer directly but found the prediction to be less stable.

4.3 Data Collection

Our pixelwise labeled dataset contains 1247 clinical head CT scans (29095 valid frames) performed from 2010-2017 on 64-detector-row CT scanners (GE, Siemens) at our affiliated hospitals. We randomly split the dataset into a trainval/test set of 934/313 stacks, called $S_{trainval}$, S_{test} respectively (S for seed).

The unlabeled set was collected using key phrase searches of radiology reports. We searched independently for positive and negative cases. The search for positive cases over 1 year yielded 1755 cases. A separate search over a shorter period identified 640 negative cases. We call this set of cases set U (for unlabeled) to be distinguished from set S. Also, 120 randomly selected cases from U (called U_{test}) were annotated at stack level in order to benchmark our system in this domain.

This dataset is smaller than the one used in chapter 3 because this work was done before the work in chapter 3.

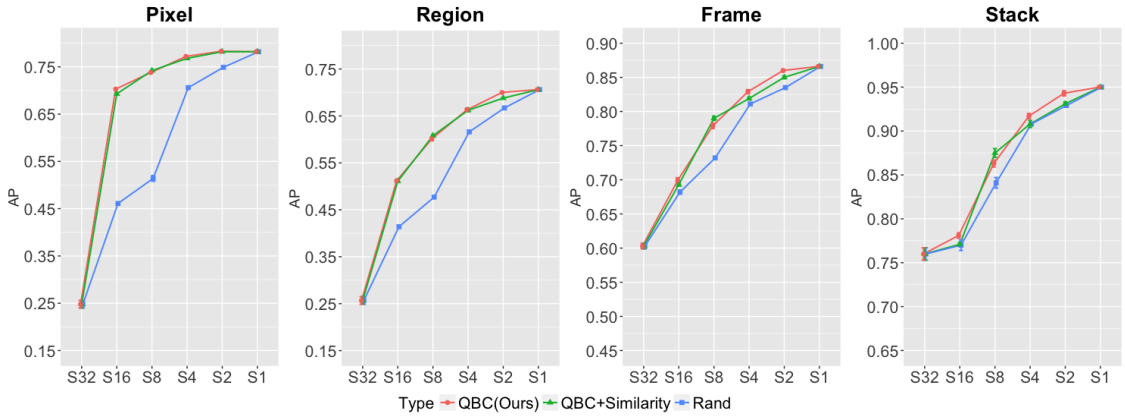

Figure 4.3: Core-set selection curves. Our system (QBC) starts to outperform [39] (QBC + Similarity) on region, frame and stack level as the dataset grows beyond one fourth of the whole set. Both QBC algorithms maintain a large gap with random baselines on pixel and region APs. For the frame and stack APs, our system still maintains a healthy margin above the random baseline for all data sizes. The region AP is computed following the definition in the first version of [16].

4.4 Experiments

Our experiments come in three parts. First, we validate the patch-based query-by-committee approach by solving the core-set selection problem on the labeled set. Then we validate the cost-sensitive approach when the seed set is already large, which is the standard setting where active learning can be useful. Lastly, we validate the entire system by actually collecting and labeling new data in the wild and showing performance gain.

Core-set Active Learning

A core-set is a subset of the training set where the empirical loss of a model is similar to that on the entire training set. In this experiment, we grow the core-set iteratively and study how the performance improves [39, 29]. For fair comparison, we strip away the cost prediction and Knapsack-solving part of our full system (See Fig. 4.1), and select examples based on their uncertainty scores alone.

We use the average precision (AP) metric to compare algorithms. Fig. 4.3 shows the performance of our query-by-committee system (QBC), suggestive annotation system (QBC + Similarity) [39], and random baseline. In this comparison, we improve [39] by using the patch-based approach for QBC + Similarity baseline, because PatchFCN [16] gives better uncertainty and similarity measures than vanilla FCN. Without it, we observed a significant performance drop. Following [39], we tried diversifying the ensemble with bootstrapping, but did not see benefit.

The experiment began with a seed set 1/32 of the training set, and doubled it by either

random sampling or active learning. In the next round, this doubled set becomes the new seed set and the process repeats. In each round, we trained an ensemble for all methods in order to compute QBC uncertainty. Fig. 4.3 shows that our system's performance at half the dataset (S2) closely matches the performance of using the whole dataset (S1) for every AP, similar to [39, 29]. However, here we use a dataset that is two orders of magnitude larger and much harder to overfit on.

Our experiment indicates that on a large dataset, QBC uncertainty alone could be sufficient to yield competitive performance, if not state-of-the-art. Without bootstrapping or pairwise similarity, our system beats the random baseline by a good margin and compares favorably with [39] in performance and time complexity. The time complexity of core-set approaches [39, 29] are dominated by the pairwise similarity computation, which is quadratic and can be expensive in practice when the seed set is too large to be grown by brute-force labeling. In contrast, our system has linear time complexity because it computes everything on-the-fly.

Cost-Sensitive Active Learning

After validating the core-set AL, we model the cost with the full system described in Fig. 4.1. We randomly select half of our labeled training set as the seed set to mimic the scenario where the seed set is large enough to render naive labeling impractical for growing the data. Yet at the same time we want the pool to be at least as large as the seed. In each iteration, we increment the data by allocating additional *time* to add labeled examples by solving the Knapsack problem. For the random baseline, we randomly select examples to add until no example can fit in the given time anymore. Fig. 4.4 shows the superiority of our system (QBC) over both uniform-cost AL (UAL) and the random baseline in such setting. The result supports Fig. 4.5 where UAL is biased toward examples with large bleeds and long labeling times. In fact, UAL selected 8/11 stacks in the first/second rounds, whereas cost-sensitive AL (CAL) selected 94/107 stacks. Due to lack of stack diversity, UAL performs worse than CAL at the stack level.

The strong gain of CAL at (+10%) not carrying over to (+20%) is explained by the ratio of unlabeled pool to the labeled training set. When the ratio is small, the data is insufficient for AL system to choose from. In Fig. 4.3, the ratio starts with 3100% and stops with 100% at S2. In Fig. 4.4, the ratio started with 100%. After (+10%) round, the ratio is 66% for CAL and 80% for Rand. The leveling off of CAL performance shows that most of the informative examples were already selected in the (+10%) round.

Active Learning in the Wild

Finally, we apply our system on the unlabeled pool described in Sec. 4.3. First, we train an ensemble on the entire labeled set. Then we select examples from the unlabeled pool under a budget of 100 hours. A neuroradiologist examined the selected cases and determined there were 115 negatives and 64 positives. There were also 51 subacute or postsurgical cases we

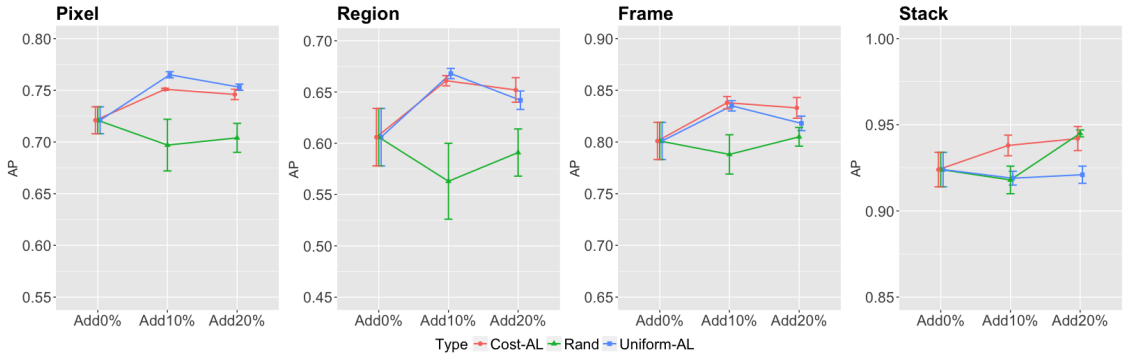

Figure 4.4: Cost-sensitive active learning. At the first iteration, the system achieves much better performance than the random baseline for all metrics. The random baseline does not improve over the seed set. In the next round, the random baseline improves the stack AP while the ALs remain the same. The error bars of AL come from the network initialization and the stochastic gradient (SGD) training. The error bars of random baseline mostly come from the random addition of data, plus the same sources of AL randomness. The time increment is 10% of the total labeling time of the pool, which simulates the situation where our budget is only a small fraction of the total labeling cost.

S_{test}	Pixel AP	Stack AP
Ens. $(S \cup U)_{train}$	$77.9 \pm 0.3\%$	$\mathbf{95.6 \pm 0.9\%}$
Ens. $S_{trainval}$	$\mathbf{78.2 \pm 0.2\%}$	$95.0 \pm 0.1\%$

U_{test}	Stack AP
Ens. $(S \cup U)_{train}$	$\mathbf{90.1 \pm 1.7\%}$
Ens. $S_{trainval}$	$85.1 \pm 0.3\%$

Table 4.1: Left: Performance on S_{test}. Compared to Ensemble $S_{trainval}$, Ensemble $(S \cup U)_{train}$ performs just as well on the pixel level and slightly outperform on the stack level. Right: Performance on U_{test}. Ensemble $(S \cup U)_{train}$ beats Ensemble $S_{trainval}$ by a good margin on the pool set.

excluded. The actual labeling time 60 hrs turned out to be within 10% of our estimate 56 hrs. We call these newly annotated examples U_{train}, to be distinguished from $S_{trainval}$ defined in Sec. 4.3. To qualitatively assess the impact of cost modeling, we show examples mined by both uniform-cost and cost-sensitive AL in Fig. 4.5.

For quantitative benchmarking, we trained an ensemble of 4 PatchFCNs from scratch with the newly augmented data (Ensemble $S_{trainval}$+U_{train}) and compared them with the ensemble trained on the original data (Ensemble $S_{trainval}$). The results on S_{test} and U_{test} are shown in Table. 4.1. We benchmark on two test sets here because we care about the performance on both seed S and pool U domains, which in practice are often not exactly the same. The gain on S_{test} shows that our method works despite the domain shift, and the strong gain on U_{test} demonstrates how a model trained on large data can be improved by collecting a little more data judiciously.

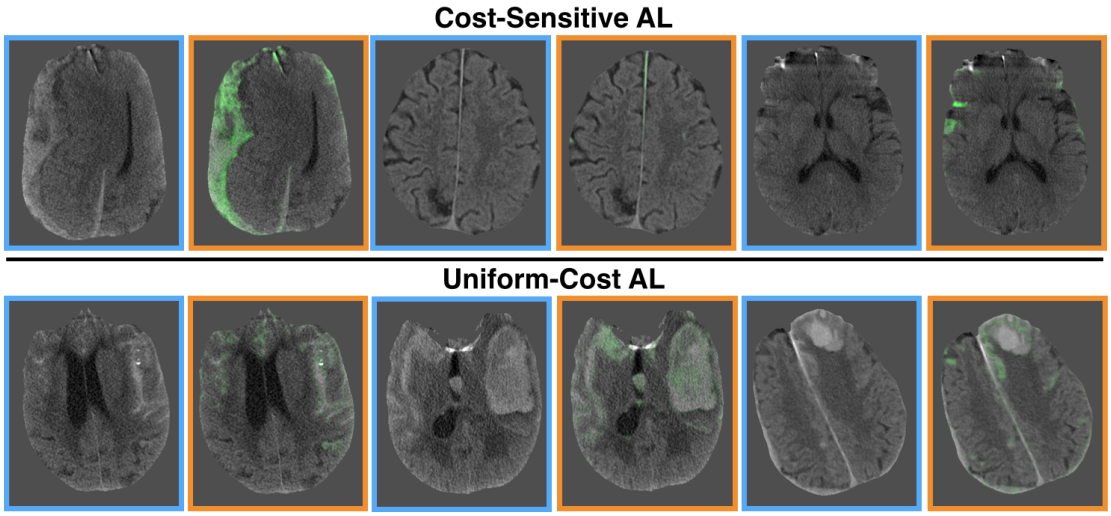

Figure 4.5: Examples selected by cost-sensitive and uniform-cost AL systems. Blue boxes are the original images, while orange boxes are the images overlaid with Jensen-Shannon divergence. The brightness of the green color indicates uncertainty. The examples selected by uniform-cost system mostly contain massive bleeds and are substantially more time-consuming for annotation, whereas examples by the cost-sensitive system are diverse and meaningful, maximizing the return on investment.

4.5 Conclusion

In this chapter, we proposed a cost-sensitive, query-by-committee active learning system for intracranial hemorrhage detection. We validated it on a substantially larger pixelwise labeled dataset than earlier works and applied it to improve the model by annotating new data from the wild. Our study demonstrates the potential of growing large medical datasets to the next level with cost-sensitive active learning.

Chapter 5

Conclusion

We study the problem of automated acute intracranial hemorrhage detection, with an application to speed up the triage at Emergency Department and reduce the workload of radiologists.

Our model PatchFCN demonstrates state-of-the-art accuracy (AUC of ROC = 0.991) on an independent test set and compares favorably to 2 out of 4 attending neuroradiologists (4-16 years of experience). We achieve this by labeling a relatively small set of head CT exams pixelwise (4.4K). PatchFCN solves the hemorrhage classification and localization problem jointly by formulating it as a semantic segmentation task. Our analyses show that it consistently outperforms the vanilla FCN byfinding a good tradeoff between the batch diversity and the amount of context.

Visualization of the PatchFCN output confirms our state-of-the-art system is able to detect cases missed by human experts and produce detailed, high quality mask on complicated cases, both of which have not been shown in the literature. In addition, we conduct an exploratory study on multiclass hemorrhage segmentation and show competitive results on both stack and pixel level with the state-of-the-art methods. Visualization confirms that PathcFCN can segment challenging multiclass cases.

Finally, we develop a cost-sensitive AL framework that aims to grow the hemorrhage dataset effectively without naive labeling. We show improved result compared to the state-of-the-art approach in core-set selection setting. Additionally, we apply the model to data in the wild and demonstrate good estimation of human annotation time as well as significant performance gain. Our framework may also be applied to other semantic segmentation tasks for which naive labeling is no longer feasible or cost-effective.

This thesis presents a series of work on hemorrhage detection. We collect the data, and demonstrate expert-level and state-of-the-art performance. Our active learning framework is a practical step to take this model to super-human level.

Bibliography

[1] Mohammad R Arbabshirani et al. "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration". In: *npj Digital Medicine* 1.1 (2018), p. 9.

[2] Babak Ehteshami Bejnordi et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". In: *Jama* 318.22 (2017), pp. 2199–2210.

[3] M Ross Bullock et al. "Surgical management of acute subdural hematomas". In: *Neurosurgery* 58.suppl_3 (2006), S2–16.

[4] PD Chang et al. "Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT". In: *American Journal of Neuroradiology* 39.9 (2018), pp. 1609–1616.

[5] Sasank Chilamkurthy et al. "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study". In: *The Lancet* 392.10162 (2018), pp. 2388–2396.

[6] Jeffrey De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". In: *Nature Medicine* 24 (Aug. 2018). DOI: 10.1038/s41591-018-0107-6.

[7] Ali Arhami Dolatabadi et al. "Interpretation of computed tomography of the head: emergency physicians versus radiologists". In: *Trauma monthly* 18.2 (2013), p. 86.

[8] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[9] Mark Everingham et al. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[10] CM Fisher, JP Kistler, and JM Davis. "Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning". In: *Neurosurgery* 6.1 (1980), pp. 1–9.

[11] Varun Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[12] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[13] Suheyla Cetin Karayumak, Marek Kubicki, and Yogesh Rathi. "Harmonizing Diffusion MRI Data Across Magnetic Field Strengths". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 116–124.

[14] Rashmi U Kothari et al. "The ABCs of measuring intracerebral hemorrhage volumes". In: *Stroke* 27.8 (1996), pp. 1304–1305.

[15] Weicheng Kuo et al. "Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 715–723.

[16] Weicheng Kuo et al. "PatchFCN for Intracranial Hemorrhage Detection". In: *CoRR* abs/1806.03265 (2018). arXiv: 1806.03265. URL: http://arxiv.org/abs/1806.03265.

[17] Jussi P Laalo et al. "Reliability of diagnosis of traumatic brain injury by computed tomography in the acute phase". In: *Journal of neurotrauma* 26.12 (2009), pp. 2169–2178.

[18] Hyunkwang Lee et al. "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets". In: *Nature Biomedical Engineering* 3.3 (2019), p. 173.

[19] David D Lewis and William A Gale. "A sequential algorithm for training text classifiers". In: *SIGIR*. 1994.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *CVPR*. 2015.

[21] Andrew IR Maas et al. "Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors". In: *Neurosurgery* 57.6 (2005), pp. 1173–1182.

[22] Dwarikanath Mahapatra et al. "Semi-supervised and active learning for automatic segmentation of crohnfffdfffdfffds disease". In: *MICCAI*. 2013.

[23] Dustin G Mark et al. "False-negative Interpretations of Cranial Computed Tomography in Aneurysmal Subarachnoid Hemorrhage". In: *Academic Emergency Medicine* 23.5 (2016), pp. 591–598.

[24] Lawrence F Marshall et al. "A new classification of head injury based on computerized tomography". In: *Journal of neurosurgery* 75.Supplement (1991), S14–S20.

[25] Luciano M Prevedello et al. "Automated critical test findings identification and online notification system using artificial intelligence in imaging". In: *Radiology* 285.3 (2017), pp. 923–931.

[26] Yao Qin et al. "Autofocus layer for semantic segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 603–611.

[27] Bob Roozenbeek, Andrew IR Maas, and David K Menon. "Changing patterns in the epidemiology of traumatic brain injury". In: *Nature Reviews Neurology* 9.4 (2013), p. 231.

[28] David S Rosen and R Loch Macdonald. "Subarachnoid hemorrhage grading scales". In: *Neurocritical care* 2.2 (2005), pp. 110–118.

[29] Ozan Sener and Silvio Savarese. "Active Learning for Convolutional Neural Networks: A Core-Set Approach". In: *ICLR*. 2018.

[30] Burr Settles. "Active learning". In: *Lectures on AI and ML* (2012).

[31] Burr Settles, Mark Craven, and Lewis Friedland. "Active learning with real annotation costs". In: *NIPS workshop on cost-sensitive learning*. 2008.

[32] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. "Query by committee". In: *Workshop on Computational learning theory*. 1992.

[33] Korsuk Sirinukunwattana et al. "Gland segmentation in colon histology images: The glas challenge contest". In: *Medical image analysis* (2017).

[34] Christopher A Taylor et al. "Traumatic brain injury-related emergency department visits, hospitalizations, and deaths – United States, 2007 and 2013". In: *MMWR Surveillance Summaries* 66.9 (2017), p. 1.

[35] AM Thabet, M Kottapally, and J Claude Hemphill III. "Management of intracerebral hemorrhage". In: *Handbook of clinical neurology*. Vol. 140. Elsevier, 2017, pp. 177–194.

[36] Joseph J Titano et al. "Automated deep-neural-network surveillance of cranial images for acute neurologic events". In: *Nat Med* 24.9 (2018), pp. 1337–1341.

[37] Katrin Tomanek. "Resource-aware annotation through active learning". In: (2010).

[38] Hongzhi Wang et al. "A multi-atlas approach to region of interest detection for medical image classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 168–176.

[39] Lin Yang et al. "Suggestive annotation: A deep active learning framework for biomedical image segmentation". In: *MICCAI*. 2017.

[40] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. "Dilated residual networks". In: *CVPR*. 2017.

[41] Yanbo Zhang and Hengyong Yu. "Convolutional neural network based metal artifact reduction in X-ray computed tomography". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1370–1381.

[42] Yishuo Zhang and Albert CS Chung. "Deep supervision with additional labels for retinal vessel segmentation task". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 83–91.

[43] Yizhe Zhang et al. "Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images". In: *BIBM*. 2016.