Deep Learning Methods for Intracranial Hemorrhage Classification

Abhilash Dhal[*], **Prashanth Duggirala**[*] **and Shawn Qiu**[*]
[*]University of California, Davis

ABSTRACT Image recognition via convolutional neural networks (CNN) has made giant strides in the last 10 years with the potential of outperforming expert-level human accuracy through adaptive learning and high dimensional feature extraction. Recently, they have gained a central stage in radiological tasks and medical diagnostics. In this project, we implement learning through a single-stage, end-to-end, convolutional neural network frameworks based on InceptionNet and EfficientNet to address the problem of classifying multiple types intracranial hemorrhages. Domain knowledge pertaining to radiology and brain hemorrhages such as "windowing" is applied for feature engineering to improve our target features for the CNN's. Overall, we achieved a mean 90% AUROC in detecting hemorrhages depending on type with single slice CTs and consumer grade hardware translating to limited computational speed and memory.

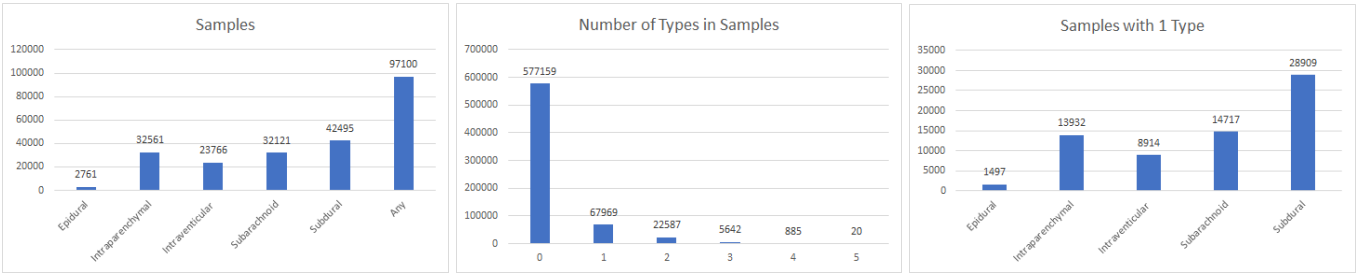
KEYWORDS Deep learning, Convolutional Neural Networks, Computed Tomography, Image Processing

Introduction

Intracranial hemorrhages cause roughly 10% of all strokes in the US, with strokes being the fifth leading cause of death. Therefore, the detection and type-identification of these conditions, is critical in timely treatment of patients, and has the potential to save many lives (RSNA 2019). While computer vision has made giant leaps since 2012, with many convolutional neural networks (CNNs) achieving expert-level results in fields such as diabetic retinopathy, skin lesions, and lymph node metastasis detection (Yamashita *et al.* 2018), intracranial hemorrhage detection today is still a manual and tedious process done by skilled radiologists. The application of these CNNs and other deep learning methods requires good data, which, until now, has not been publicly available for intracranial CT scans.

The Radiological Society of North America (RSNA), since 2017, has put out "AI Challenges" by publicly releasing anonymized CT scans of various body parts with associated labels. In 2019, RSNA put out an AI Challenge to detect intracranial hemorrhages alongside nearly 700,000 labeled samples of cranial CT scans. This data set is absolutely massive compared to the data sets used by previous studies even though it only contains 2D slices. For example, RADnet used roughly 300 samples (Grewal *et al.* 2017) and PatchFCN used 4300 samples (Kuo *et al.* 2019), though each sample contains a stack of roughly 32 slices.

PatchFCN, a recently released fully convolutional neural network, has showed promising results in using deep learning for segmentation of acute intracranial hemorrhages in CT scans using targeted labelling and annotations from field experts. However, they do not classify the type of hemorrhage and use a stack of 27-38 CT slices for each subject. Our approach is more individual in the sense that we only information from one slice at a time to classify if there is a hemorrhage and if there is one, what type it is.



(a) (b) (c)

Figure 1 Distribution of sample images: (a) All samples (b) Types per Sample (c) Sample with 1 Class

Methods

Class Balancing on input data

The data set provided by RSNA, while large and carefully labelled, poses a number of challenges. First, we do not have the computational resources at our disposal to efficiently train on the full data set. Second, the data set comes with many biases that can skew the results of models trained on the full data set.

- Less than one-seventh of samples provided have hemorrhages: this will skew predictions towards under-predicting the occurrences of hemorrhages if not accounted for.
- The samples contain labels for 5 different categories of hemorrhages Fig 1 (a): Epidural, Intraparenchymal, Intraventricular, Subarachnoid, and Subdural. These types correspond to different locations and types of tissues where the bleeding occurs: this will require a model that handles multiple classes.
- Many of the samples simultaneously have multiple types of hemorrhages, including 20 that have all 5 types Fig 1 (b): we need to consider whether this will affect the training of our model.
- The distribution of samples with each of the 5 categories is not even (e.g there are significantly fewer samples with Epidural hemorrhages than any other type) Fig 1 (c): this distribution may or may not reflect the prevalence of each class in the field. In order to maximize generalizability, this needs to be accounted for.

With these issues and our constraints in mind, random majority under-sampling with multiple classes was used. 1000 samples of each category were randomly selected. Of these samples, each sample only contains the type of hemorrhage for which it was chosen. Additionally, 1000 samples with no hemorrhages were selected for comparison against these samples for a total of 6,000 samples. This forms a relatively small data set compared to the full set, but one that is free from observable biases - this make training the model straightforward.

Windowing CT Scans

Head CT plays a critical role in the evaluation of intracranial abnormalities (such as trauma, stroke, and hemorrhage). CT images are generated using X-ray beams. The amount of X-rays absorbed by tissues at each location in the body is mapped to Hounsfield units (HU). The denser the tissue, the more the X-rays are attenuated, and the higher the number of HUs. Water is always set to be 0 HU, while air is 1000HU, and bones have values between several hundred to several thousand HU. (Xue Z *et al.* 2012)

Windowing, also known as grey-level mapping, is the process in which the CT image are manipulated; doing this will change the appearance of the picture to highlight particular structures.

This allows different features of tissues to be seen and enables viewers to focus on certain tissue of interest by maximizing subtle differences among the tissues. Windowing is controlled by two parameters: window level (WL) and window width (WW). The window width (WW) as the name suggests is the measure of the range of HU values that an image contains. The window level (WL), often also referred to as window center, is the midpoint of the range of the HU values displayed. The brightness of the image is adjusted via the window level. The contrast is adjusted via the window width. As illustrated in Figure 2, only the tissues with HU values within the specified window ([WL-WW/2, WL+WW/2]) are mapped onto the full range of gray scale; the tissues with HU values above (>WL+WW/2) or below window (<WL-WW/2) are set to be all white or all black.

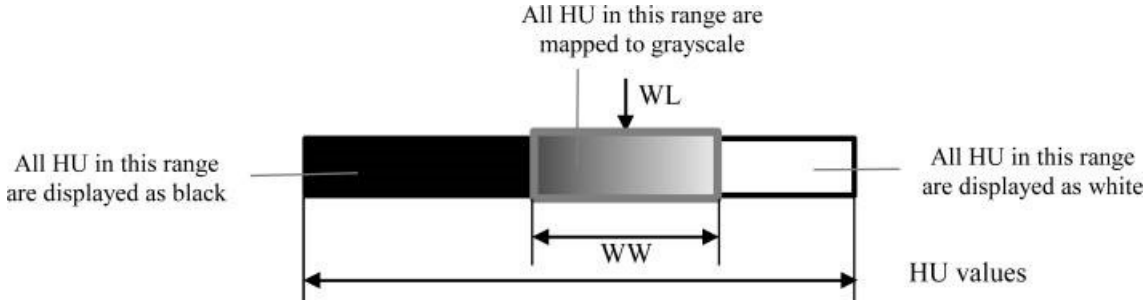


Figure 2 Example Window

Pre-processing the inputs

The samples provided are stored in the DICOM format. While this is the industry standard format for medical images, it is unfamiliar to most computer scientists. The DICOM format provides attributes to store key values alongside a 512x512 pixel_array. Some of these attributes include patient ID, study ID, the window set by the radiologist, etc. The pixel values in the pixel_array are also not in standard RGB format but in Hounsfield units.

DICOM image data, due to being stored as Hounsfield units, has a range of values much wider than the typical png or jpg image. Let's remind ourselves what the scans look like if we include the full range of values it looks like fig. 3(a). They are not very useful for detecting hemorrhages. The DICOM images come with metadata specifying a window center and width. We use these values for windowing to produce an image as shown in fig. 3(b), these metadata values are used to visualize the scan in the range of values that correspond with brain matter.

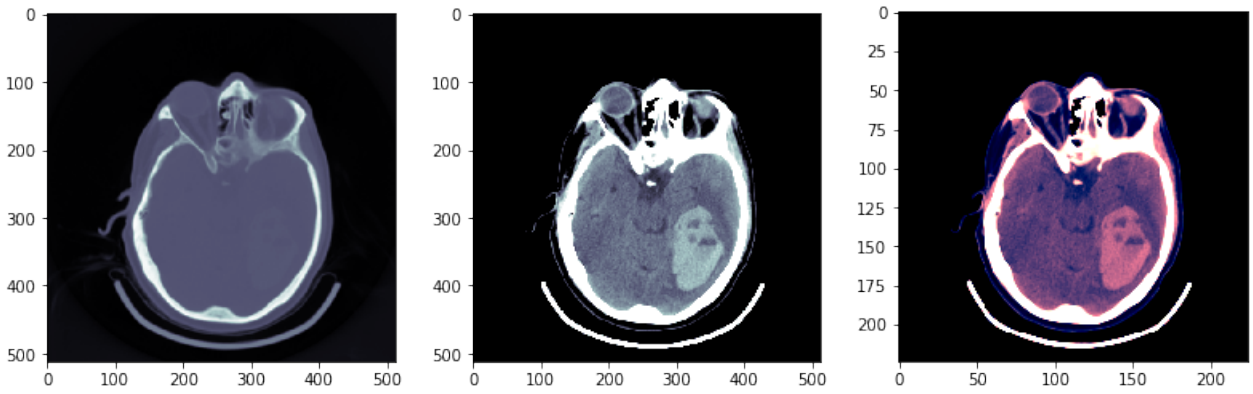
There are at least 5 windows that a radiologist goes through for each scan:

- Brain Matter window : W:80 L:40
- Blood/subdural window: W:130-300 L:50-100
- Soft tissue window: W:350–400 L:20–60
- Bone window: W:2800 L:600
- Grey-white differentiation window: W:8 L:32 or W:40 L:40

We built on the hypothesis that different hemorrhages become more obvious at different window settings so we can include more than one window in our training image by storing a different window in each of the three channels, we used three types of windows corresponding to brain, blood/subdural and soft tissues, see fig. 3(c).

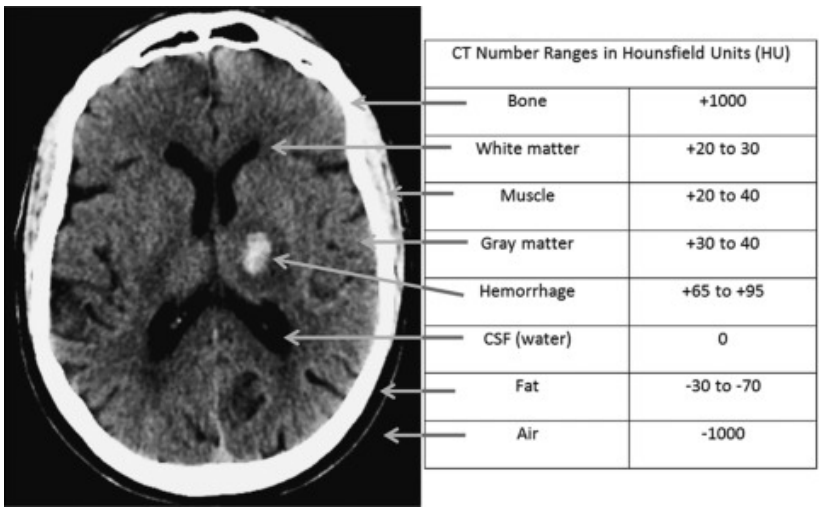
Motivation for Convolutional Neural Networks application

Initially, we performed filtering similar to thresholding on the windowed image to highlight only the hemorrhage based on the Hounsfield value of blood (i.e. 40 to 65). Shown below in fig. 5 is an



(a) (b) (c)

Figure 3 (a) Unprocessed image (b) Metadata based windowing (c) Our windowing



transfer

Figure 4 Hounsfield units for various types of tissue

example of a thresholded image in comparison to the original image. We tried this method to see how close we could come to solving the problem with traditional image processing (Thresholding, opening/closing, contouring) + classical machine learning (SVM and Random Forest) approaches but as expected, we didn't achieved favourable results so we decided to use the deep learning approach.

Therefore, we moved onto models based on convolutional neural networks (CNN). A CNN is a class of prediction models developed in the domain of computer vision for learning spatial features across high dimensional tensors and mapping them them into low dimensional labels. The simplest CNN architecture comprises of a convolution, pooling and dense layer to perform feature extraction followed by label mapping. A major application of this toolkit is in the domain of radiology images (Yasaka *et al.* 2017). On account of the high dimensional complexity involving CNN's and limited computational resources (thus limited number of samples), our study was motivated to implement a pre-trained model in order to avoid the compute time and memory involved in training the initial filters that are common to most image recognition tasks such as edge detection.

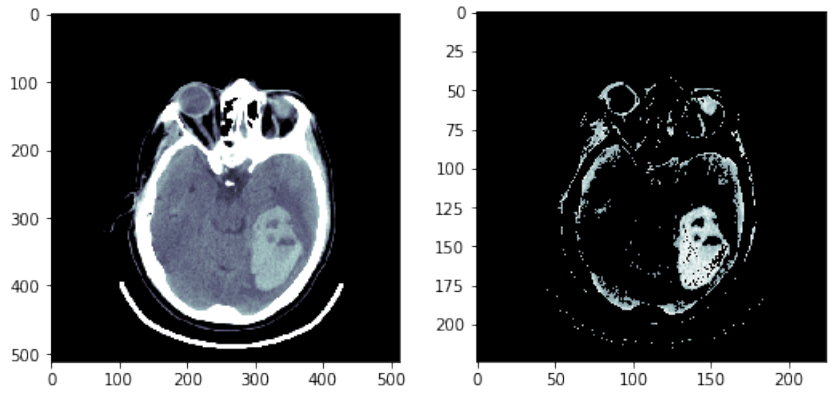


Figure 5 Comparing the (a) windowed image and (b) thresholded image

Binary classification of CT images: Hemorrhage or no Hemorrhage

InceptionV3 (Szegedy *et al.* 2015) was chosen due to having the the richest feature set representation across the widest range of images today. We perform fine tuning on the pre-trained weights from ImageNet (Krizhevsky *et al.* 2012). On top of the network, the following layers were added: (1) 1 global average pooling layer (2) 3 dense layers with relu activation and 50% dropout and (3) a final prediction layer of size 2 with softmax activation. Binary cross-entropy was used as the loss function.

The data set for each class (1000 samples of a hemorrhage and 1000 samples of no hemorrhages) was further divided into training and validation sets with a 80:20 split. The training was done with a batch size of 32 and the Adam optimizer with a learning rate of 0.001, a beta_1 of 0.9, and a beta_2 of 0.999 for 20 epochs due to our limited computational power. The learning rate is reduced by the order of 10 every time there is a plateau in the loss function.

Multi-class classification

We then further created multi-class models based on EfficientNet (Tan and Le 2019) and NasNet (Zoph *et al.* 2017). We added the same additional layers as our InceptionV3-based model, but with a final prediction layer of size 6 and a loss function of categorical cross-entropy to gain multi-class functionality.

The data set used was our full curated 6,000 sample set which was then divided into training and validation sets with the same 80:20 split as our binary-classification model. The training also done on the same set of hyper-parameters. Training took roughly 3 hours on the free resources provided by Google Colab, equivalent to no more than a Nvidia Geforce GTX1070.

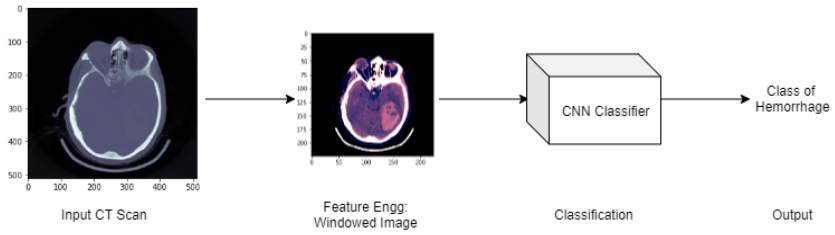


Figure 6 Prediction Pipeline

Model Validation

In our study, the validation of deep neural nets InceptionV3, EfficientNet and NasNet performance is done by the area under the curve (AUC) for the receiver operator characteristics graph. AUC measures the predictive power of our models, which means it quantifies the controlling of false positives in our detection. We first obtain our prediction probabilities from the trained models for each of the classes and then construct the ROC curve by plotting the true positive rate against the false positive rate for each class. Then the y axis value and x axis value for each threshold point of the ROC curves are calculated as:

$$(TPR)_i = TP_i/(TP_i + FN_i) \tag{1}$$

$$(FPR)_i = FP_i/(FP_i + TN_i) \tag{2}$$

TP_i, FP_i, TN_i and FN_i are the number of true positives, false positives, true negatives and false negatives for threshold of i where $i \in (0, 1)$.

Results

Binary Classification of hemorrhages

The binary classification model performs with an accuracy of 85.5% for one partition of training-validation set but we did not perform cross-validation because of time and compute constraints. Instead we use the AUC for validating our performance as it is a more stable measure as compared to validation accuracy when our data set is fixed for training-validation cycle. Therefore, we report performances for models in terms of AUC values only. For the binary classification, we obtain an AUC of 0.92 (figure 7) with our InceptionV3 net model.

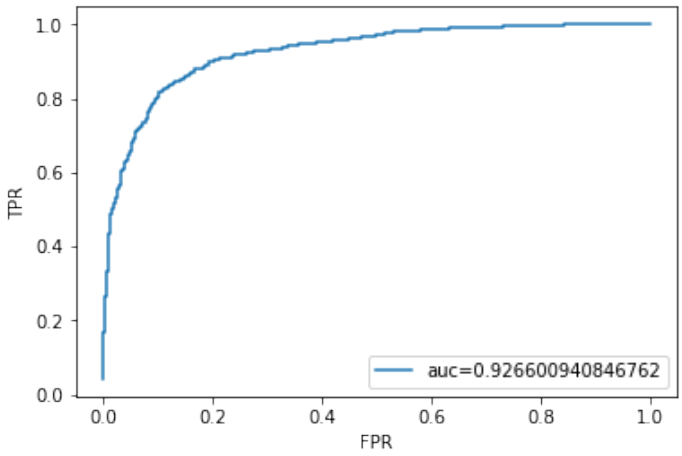


Figure 7 Receiver operator characteristics with InceptionV3

Multi-class classification of hemorrhages

The results from our multi-class EfficientNet model is significantly better than the results from the InceptionV3-based model and NasNet based model. The AUC results for each class in this model including no hemorrhage are shown in Table 1 for EfficientNet and NasNet. We additionally add the notation of RSNA0 through RSNA 5 to each class for further tables and figures. In just 25 epochs, we achieved AUCs averaging 88% across our 6 classes (figure 8). This average is dragged down by the 81% and 83% AUC results from Subarachnoid and Subdural classes.

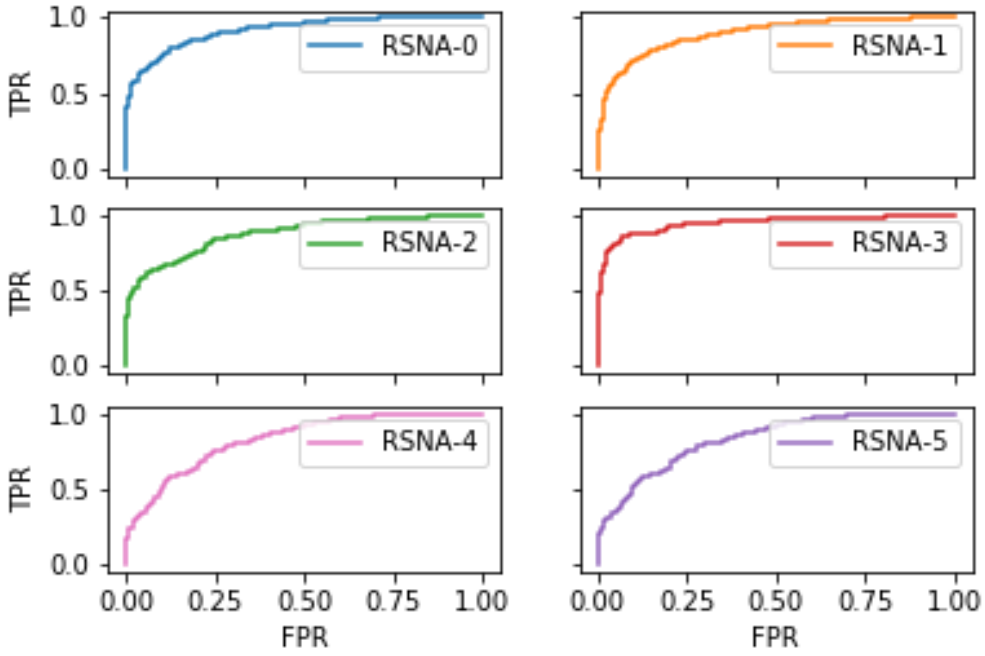


Figure 8 Receiver operator characteristics with EffNet using relu activation

Class	Notation	EffNet-Relu	EffNet-Tanh	NasNet
No Hemorrhage	RSNA0	0.917	0.916	0.839
Epidural	RSNA1	0.893	0.892	0.816
Intraparenchymal	RSNA2	0.881	0.896	0.775
Intraventricular	RSNA3	0.949	0.946	0.883
Subarachnoid	RSNA4	0.832	0.829	0.70
Subdural	RSNA5	0.811	0.818	0.72

Table 1 AUC value for NasNet Effnet tanh, relu models

Discussion

We are pleased with the results (90% mean AUC) achieved by our multi-class model. However, there were areas in our model that were scaled back in order to run in a reasonable time-frame. The following features of our model were chosen to help minimize training time with limited computational power. With more computational resources, we would make the associated changes to improve our AUC:

- Adam optimizer: Adam was chosen due to it's computationally efficient nature, even though stochastic gradient descent (SGD) could yield better results as it is guaranteed to find the global minima.
- Number of epochs: the models were limited to 25 epochs. More epochs can lead to better results. For example: PatchFCN used 400 epochs in their training.
- Reduced data set: The full RSNA data set contains significantly more samples than we were able to use. In some classes, the full data set includes up to 30x more samples than we used. This data would be particularly useful in the Subarachnoid and Subdural classes where we had our worst results.

In the three model architectures that we implemented, the total number of trainable parameters, type of convolutions, model scaling and optimizing search spaces are some of the important factors influencing the performance of our models. Due to the complexity of their architecture, we see multiple basic architectures of Recurrent neural networks(RNN's), Reinforcement learning(RL), CNN's and other components combined into a large model in the three neural networks that we implemented. The motivation for such large scale neural networks is focused on addressing generalization of the model's classification. Model scaling, compound model scaling and Search Space optimization are implemented in InceptionNet, EffNet and NasNet (Zoph *et al.* 2017) respectively. Compound scaling in Effnet uniformly scales the height, width and resolution improves on single dimensional model scaling procedures for InceptionNet. In addition, the compound scaling procedure is also adaptive to the model architecture as it does not perform arbitrary scaling. Compound scaling therefore improves overall performance while optimizing the architecture adaptively. NasNet in contrast performs search space optimization techniques that look for local minima for trainable parameters as compared to global minima, which not necessarily will improve the performance of classification.

We also tried a different approach for classification using conditional adversarial networks (Frid-Adar *et al.* 2018). We conjectured that training our data comprising of real input data distributions and generated data distributions could be used as input to the discriminator for classifying the labels in conditional generative adversarial networks(GAN's). conditional GAN's generate probabilities for output classes for the real image which can be used for classification . The training time for generating good fake distributions in terms of the convergence of conditional GAN's is very high and therefore we do not explore CGAN's.

In conclusion, this study set out to detect intracranial hemorrhages with the images provided by the 2019 RSNA AI Challenge. We were able to achieve a roughly 90% prediction power depending on the type of hemorrhage by using convolutional neural networks. Given the limited resources used, this is a great result compared to the minimum 82.5% accuracy rate required of human radiologists and compared to the 82% accuracy achieved by RADnet. While our result still falls short of the astonishing results of PatchFCN, which achieves a 99% accuracy rate, we were able to do this with only roughly 1/32nd the data per sample used in all previous works that we are aware of - our approach is the only one that uses only a single slice of a CT scan instead of a 3D volumetric stack of 27-38 image slices and classifies the type of hemorrhage. With more computational resources, we hope that our approach can help doctors improve patient outcomes in an applicable way.

Literature Cited

Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger, *et al.*, 2018 Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. CoRR **abs/1803.01229**.

Grewal, M., M. M. Srivastava, and S. Kumar, Pulkit andVaradarajan, 2017 Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans **arXiv:1710.04934**.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012 Imagenet classification with deep convolutional neural networks pp. 1097–1105.

Kuo, W., C. Hne, P. Mukherjee, J. Malik, and E. L. Yuh, 2019 Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proceedings of the National Academy of Sciences **116**: 22737–22745.

RSNA, 2019 Rsna intracranial hemorrhage detection. https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection, Accessed: 2019-10-27.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 2015 Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567**.

Tan, M. and Q. Le, 2019 EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, Long Beach, California, USA, PMLR.

Xue Z, A. S., L. LR, D.-F. D, and T. GR., 2012 Window classification of brain ct images in biomedical articles. AMIA Annual Symposium proceedings. AMIA Symposium, 2012, 1023–1029. **135**: 1023–1029.

Yamashita, R., M. Nishio, R. K. G. Do, and K. Togashi, 2018 Convolutional neural networks: an overview and application in radiology. Insights into Imaging **9**: 611–629.

Yasaka, K., H. Akai, O. Abe, and S. Kiryu, 2017 Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: A preliminary study. Radiology **286**: 887–896.

Zoph, B., V. Vasudevan, J. Shlens, and Q. V. Le, 2017 Learning transferable architectures for scalable image recognition. CoRR **abs/1707.07012**.

Author Contributions

Authors agree that everyone contributed roughly equally in different ways

Abhilash Dhal:

- Related works research
- Feature Engineering
- Windowing and thresholding
- Binary Classification with GAN's
- Multi-class models (EfficientNet and NasNet) built on base model
- Various experiments on learning (SGD, different data sets, epochs, dropout rates, base-model used, etc)
- Exploring performance measures like AUC(ROC), Confusion matrix, Precision and Recall.

Prashanth Duggirala:

- Related works research
- Classical methods experiments
- Base (starter) Models for Binary and Multi-class models
- Windowing
- Feature Engineering
- Image Data Generator
- Binary classifier
- Various experiments on learning (SGD, different data sets, epochs, dropout rates, base-model used, etc)
- Exploring performance

Shawn Qiu:

- Initial draft of report
- Data Balancing
- Feature Engineering
- Windowing and thresholding
- Various experiments on learning (SGD, different data sets, epochs, dropout rates, base-model used, etc)
- Related works research
- Project management