Intracranial Hemorrhage Detection in CT Scans using Deep Learning

Tomasz Lewicki, Meera Kumar, Raymond Hong, Wencen Wu*

Computer Engineering Department
San José State University
San José, CA, USA
Email: tomasz.lewicki@sjsu.edu,meera.kumar@sjsu.edu,raym.hong@gmail.com,wencen.wu@sjsu.edu

***Abstract*—In intracranial hemorrhage treatment patient mortality depends on prompt diagnosis based on a radiologist's assessment of CT scans. In this paper, we investigate the intracranial hemorrhage detection problem and built a deep learning model to accelerate the time used to identify the hemorrhages. To assist with this process, a deep learning model can be used to accelerate the time it takes to identify them. In particular, we built a convolutional neural network based on ResNet for the classification. Using 752,803 DICOM files collected from four international universities by the Radiological Society of North America (RSNA) [1], we trained and tested a ResNet-50 based model for predicting the hemorrhage type. Our model has an accuracy of 93.3% in making the correct multiclass prediction and an average per-class recall score of 76%. We show it is possible to achieve an average recall of 86% while maintaining 70% precision via tuning the prediction thresholds. Lastly, we show real-world applicability by deploying a simple web application. The source code for training, metrics evaluation and web application is available at [2].**

***Index Terms*—Intracranial hemorrhage, head computed tomography, feature recognition, deep learning**

I. Introduction

Intracranial hemorrhage (ICH) affects various patients who suffer from trauma, stroke, aneurysm, vascular malformations, high blood pressure, illicit drugs and blood clotting disorders. Different types of ICH can be identified by their distinct shape, location, and size. Early diagnosis of ICH, preferably within 24 hours, is key in decreasing patient mortality. Diagnosis time includes taking a head computerized tomography (CT) and a radiologist making a diagnosis based on CT interpretation. CT scans consist of a series of X-ray images taken in different angles, combined to form 3D cross-sectional images of the blood, bone, and soft tissue within the body.

Advances in image recognition using machine learning, has increased the applicability of ML models such as neural networks for medical image processing. Deep learning systems recognize meaningful patterns and features within large datasets without explicit directions. As such, they can be trained from end to end. Utilizing machine learning to infer and perform complex cognitive tasks such as analyzing CT scans that normally requires an expert, radiologist.

In this study, we use a deep convolutional neural network for detection and classification of different ICH on unenhanced CT scans. Fast detection of ICH and differentiating it from types of strokes and other brain disease can prompt appropriate treatment and mitigate lasting brain damage and mortality.

The rest of the paper is organized as follows. Section II introduces some related works. Section III presents the dataset we use in this study. The data pre-processing, CNN model, and training are introduced in Section IV and results are presented in Section V. Section VII concludes the paper and discusses some future work.

II. Related Works

Convolutional neural networks (CNN) are widely used in computer vision. CNNs are capable of surpassing human level performance in image classification [3]. While these tests have been done on generic classification of the ImageNet dataset [4], there has been work towards validating machine learning results with human specialists for more complex identification [5], [6]. Chilamkurthy et. al. constructed their own dataset by collecting head trauma related CT scans from local hospitals to train their algorithms. Their algorithms' performance when compared with three radiologists came out higher. Though these tasks are not considered "solved", they provide reliability for using neural networks to aid with medical image analysis.

Groups with the best performance stand out through novel processing techniques on the data [4]. A typical CT scan results in multiple DCM files that contain an image of one cross section. Data can be processed one image at a time (2 dimensional) or in a series to maintain the 3 dimensionality. A study on diagnosing Alzheimer's disease compared both 3D and 2D preprocessing of their dataset [7]. The 2D version of their CNN outperformed the 3D, but their best results came from a consolidation of both networks.

Models for processing medical images benefit from pre-training on general datasets such as ImageNet. ResNet-50 has been used in a few brain image analysis projects, namely for Alzheimer's disease [8] with nearly perfect multi-class prediction accuracy. Another study compared various DL models to detect brain abnormalities to find RestNet-50 resulted in the best classification accuracy [9]. There studies were done on MRI scans. Applying DL models to CT scans is still under-investigated.

978-1-7281-7022-0/20/$31.00 ©2020 IEEE
DOI 10.1109/BigDataService49289.2020.00033

III. Dataset

Our dataset, provided by the Radiological Society of North America (RSNA) [1] with patient files from 4 international hospitals, contains a total of 752,803 labeled DCM files containing cross-sectional images of the brain that are 512x512 in size. We set aside 10% of randomly selected examples as the test set. Each file is labeled with one or more types of hemorrhage, or no hemorrhage at all. Fig. 1 shows what type of hemorrhages could be present in our dataset. Labels are given in the form of a CSV file that contains six lines for each patient ID, with each line corresponding to a type of hemorrhage, followed by a boolean value of whether that type of hemorrhage is present in the image.

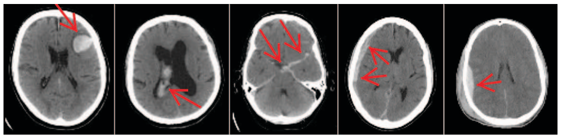

Fig. 1. Types of hemorrhages: (From Left to Right) Intraparenchymal, Intraventricular, Subarachnoid, Subdural, Epidural. Source: [1]

After reformatting, the 6 patient labels are condensed to a N-hot encoded 1x6 array, where 5 elements match specific hemorrhage types and one element denotes ”any” class. Table I shows a small sample of how the labels are represented after preprocessing. It is important to note that these types are not mutually exclusive. In other words, a single image can contain more than one type of hemorrhage. There are 85.7% negative samples (no hemorrhage at all), 10.1% samples with one diagnosis and 4.2% positive samples with more than one diagnosis.

TABLE I
Small sample of patient data and their corresponding labels

image id	any	epid.	intrap.	intrav.	subar.	subd.
ffff82e46	0	0	0	0	0	0
ffff922b9	1	0	0	1	0	0
ffffb670a	1	0	0	0	1	0

The dataset contains heavy negative bias (most of the scans contain no type of hemorrhage at all), as well as an imbalance between positive classes, as depicted in Fig. 2. These two properties of the dataset are addressed by class weights in the loss function and recall/precision tuning respectively and are discussed in section IV.

IV. Methodology

A. Data Pre-processing

CT scans are generated using X-rays. The denser the tissue, the more X-rays are attenuated and the higher the resulting pixel intensity. For feature extraction, windowing maximizes the subtle difference between features [10]. Using the documented range of window level (WL) and window width (WW) for the desired tissues, only tissues with the desired Hounsfield units (HU) are mapped into the three channels of the input tensor, as visualized in Fig. 3. Each window's values are normalized to ([0,1]). As we are looking for bleeding within the brain, subdoral/blood (WL=80, WW=200), and brain matter (WL=40, WW=80) windows were chosen. Though both provide information on soft tissues their range of HU is different. We chose bone (WL=600, WW=2800) for our third feature in order to identify abnormalities that occur by the skull, namely in subdural and epidural cases.

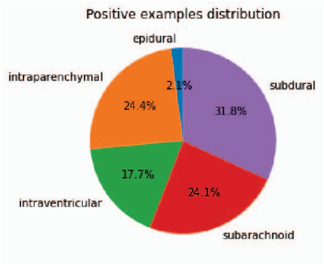

Fig. 2. Pie chart depicting the distribution of hemorrhage classes in the dataset

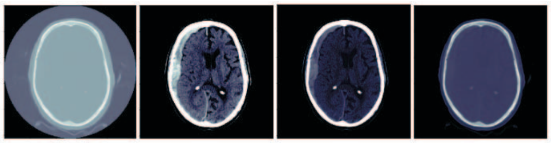

Fig. 3. Preprocessing (From Left to Right): Uncontrasted Scan, Brain Window, Subdural/Blood Window, Bone Window

B. Model

1) CNN architecture: We use the convolutional base from ResNet50 [11] as a feature extractor, onto which we stack a 2 layer fully-connected classifier *(refered to as the ”head” in the rest of this paper)*. The head consists of:

- 256 neuron hidden layer with ReLU activation
- 6 neuron output layer with sigmoid activation

Although there have been architectures outperforming it classification tasks, we chose ResNet for its simplicity and regularity of structure. Fig. 4 illustrates the data flow during model inference.

The input tensor dimensions are (224,224,3) with values normalized within the range $[0, 1]$ in the windowing process. The intermediate feature vector (the output from the conv. base) dimensions are: (7,7,2048) (2048 channels of 7x7 matrices). The output vector of the model is a 6-dimensional vector, with 5 values corresponding to the likelihood of diagnosis for each type of hemorrhage. The 6th value corresponds to the likelihood of ”any” class. Even though this label can be explicitly calculated as a function of the 5 remaining labels (i.e. a 5-input OR gate), we deliberately choose to include it in training in order to later compare the performance between the learned and explicit approach to computing this label.

2) Task and loss function definition: As we showed in fig. 2, there is a significant number of positive examples with more than one diagnosis. This determines the task of predicting such

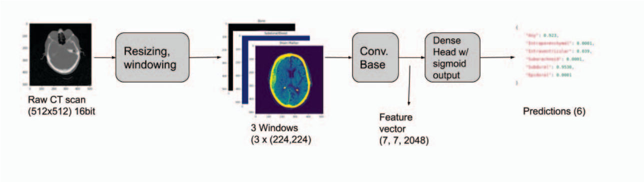

Fig. 4. Data flow during model inference.

diagnoses as Multi-Label Classification. The loss function is defined as follows:

$$loss = -\frac{1}{C}\sum_{i=1}^{C} w_i \cdot [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (1)$$

Where: y_i - ground truth for i-th class, p_i - prediction for i-th class, C - number of classes w - weight vector

Weighted cross-entropy loss compensates for the class imbalance in the dataset using the weight vector. For simplicity, we defined the weights vector as $w = [1.0, 2.0, 1.0, 1.0, 1.0, 1.0]$ (2.0 value corresponds to the epidural hemorrhage, which is the rarest). These values could also be determined via hyperparameter optimization.

C. Training

The parameters of the convolutional base are initialized with weights pre-trained on the ImageNet dataset. Even though ImageNet images are vastly different from CT scans, in an early experiment we were able to achieve 90% accuracy and 32% recall by simply freezing the convolutional feature selector pre-trained on ImageNet, and only training the dense head. Thus we concluded that ImageNet weights will serve as a good weight initialization for the conv. base. The fully-connected head is initialized with random weights sampled from normal distribution. We use Adam [12] optimizer with learning rate of $1 \cdot 10^{-5}$, determined via hyperparameter optimization. We train the network in mini-batches of 16 examples. The training takes 10 epochs on Nvidia Tesla P100 accelerator, which corresponds to approximately 48 hours.

V. Results

A. Key metrics

For model evaluation, we randomly sample 10% of the entire dataset (approximately 75k examples) as a test set. Our model yields 98.3% accuracy and 93.3% per-image accuracy (i.e. the model predicts all classes correctly in 93.3% cases). However, due to the high dataset negative bias, accuracy would be a rather poor standalone performance metric. Consequently, we keep track of per-class recall and per-class precision for more comprehensive evaluation.

TABLE II
Per-class performance metrics

Metric	any	epid.	intrap.	intrav.	subar.	subd.
accuracy	0.965	0.998	0.985	0.99	0.979	0.978
precision	0.914	0.913	0.900	0.864	0.811	0.905
recall	0.833	0.682	0.765	0.846	0.716	0.722

Recall is our most important measure of success, as it corresponds to the number of sick patients with the right diagnosis out of all the labeled sick patients. False negatives are particularly dangerous for the patient and must be avoided in medical diagnosis. Average recall is 0.76, with the lowest recall for epidural hemorrhage (0.682) and the highest recall for intraventricular hemorrhage (0.846). The class with the fewest examples has the lowest recall, which implies the class weight vector may be further optimized. Additionally, the relatively high precision scores imply that there is potential to trade off precision for higher recall.

B. ”any” label - predicted or explicit?

In section IV-B1 we prompted the two approaches to compute “any” label – either to learn it along with other labels, or to compute it as a boolean ’OR’ of the other 5 predictions. As shown in the Table III, the former method performs better in terms of recall by $2.5p.p.$ and thus we decide to use the model-based predictions. Such a result is rather surprising – even though we know the dependency explicitly (5-input OR function) it is better to derive the result directly from the neural network. A possible explanation, is that there exist features which the classifier for ’any’ class was able to learn and the individual classifiers omitted.

TABLE III
Comparison of the two approaches to ”any” prediction

Metric	“any” (predicted)	“any” (explicit)
accuracy	0.965	0.964
recall	0.833	0.809
precision	0.914	0.927

C. Precision-recall trade-off and ROC

One of the ways to tune the predictive ability of our model is to adjust the threshold for the neutral network’s binary predictions. Threshold manipulation affects the precision-recall trade-off, as shown by the P-R curves for each class in fig. 5.

Table IV shows the precision, recall and accuracy values yielded from predicting “any” label on the test set at different threshold values. The table V shows the results obtained with various threshold that keep a fixed minimum precision of 0.7. Such a tuned model, yielding higher recall but allowing for a certain, limited loss of precision (thus giving more false positive diagnoses) could be significantly more useful to the end user. However, the decision how big a precision loss is acceptable for a given class of hemorrhages should be based on expert knowledge and solid clinical evidence, rather than any technical justification.

We also performed a ROC *(Receiver Operating Characteristic)* analysis on the test set. Area under the resulting ROC curve is 0.985. The detailed ROC curve figures are available in the source code.

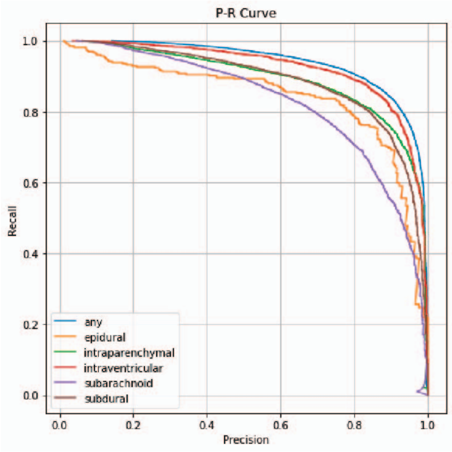

Fig. 5. Per-class P-R curves

TABLE IV
"ANY" CLASS RECALL-PRECISION TRADEOFF

Threshold	precision	recall	accuracy
0.0	0.5	0.973	0.595
0.001	0.55	0.966	0.89
0.002	0.6	0.959	0.906
0.004	0.65	0.95	0.921
0.008	0.7	0.939	0.934
0.018	0.75	0.925	0.945
0.045	0.8	0.906	0.954
0.122	0.85	0.877	0.96
0.345	0.9	0.839	0.964
0.828	0.95	0.763	0.961

VI. DEPLOYMENT CONSIDERATIONS

Through a simple web application, we show the real-world applicability of our model. User is prompted to choose a raw DCM file, and given the likelihoods of hemorrhage classes in response. The full graph forward pass takes 366ms on a laptop computer with Intel i7-5600U CPU and consumes 2.5GB of RAM. This shows, that such a system could be deployed on a mid-range computer in a medical care facility, even without dedicated hardware accelerators. Additionally, local deployment mitigates the the need to upload patient's data to the cloud, which would require compliance with strict health privacy regulations.

VII. CONCLUSIONS AND FUTURE WORK

We built a model that can analyze single frame CT scans with comparative accuracy to a radiologist. Our model was trained on the RSNA dataset [1], a large labeled dataset of 752,803 unenhanced CT scans. Using windowing (contrast enhancement) for preprocessing, the CT scans are converted to 16 bit dicom images into 3 channels with matrices of floating-point numbers normalized in (0, 1) range. To train our model with limited bias for the disproportionate independent hemorrhage classes, a weighted logistic loss function was used for multiclass classification. The model has per-class accuracy of 98% and per-image-accuracy of 93.3%. We report recall with an average of $.875 \pm 0.053$ with precision set to 0.7.

Future improvements include evaluating the system with different convolutional base architectures, such as deeper versions of ResNet (ResNet-101 and ResNet-152). Another enhancement would be altering the loss function by introducing different weights for positive and negative examples. Such a doubly-weighted loss function would allow to address the precision/recall tradeoff during training, possibly with better results. Another great addition to the system would be extending it with federated learning capability, where the client nodes can perform backpropagation on edge. This way, the system could scale incrementally, and also preserve the sensitive patient's information within the medical facility.

TABLE V
PER-CLASS METRICS YIELDED WITH A FIXED PRECISION MARGIN.

class	any	epidur.	intrap.	intrav.	subara.	subdur.
precision	0.7	0.7	0.7	0.7	0.7	0.7
threshold	0.008	0.105	0.035	0.088	0.125	0.034
recall	0.939	0.84	0.879	0.922	0.794	0.875
accuracy	0.934	0.998	0.976	0.984	0.975	0.969

REFERENCES

[1] Kaggle and RSNA, "Intracranial hemorrhage detection dataset," 2019. [Online]. Available: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/data

[2] T. Lewicki, https://github.com/tomek-l/brain, [Online].

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015. [Online]. Available: http://arxiv.org/abs/1502.01852

[4] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[5] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *CoRR*, vol. abs/1707.01836, 2017. [Online]. Available: http://arxiv.org/abs/1707.01836

[6] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. Campeau, V. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study," *The Lancet*, vol. 392, 10 2018.

[7] X. W. Gao, R. Hui, and Z. Tian, "Classification of CT brain images based on deep learning networks," *Computer Methods and Programs in Biomedicine*, vol. 138, pp. 49–56, 2017. [Online]. Available: https://doi.org/10.1016/j.cmpb.2016.10.007

[8] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, and C. P. Fulton, "Classification of alzheimer's disease with and without imagery using gradient boosted machines and resnet-50," *Brain Sciences*, vol. 9, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2076-3425/9/9/212

[9] M. Talo, Ö. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya, "Convolutional neural networks for multi-class brain disease detection using MRI images," *Comp. Med. Imag. and Graph.*, vol. 78, 2019. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2019.101673

[10] J. E. Barnes, "Characteristics and control of contrast in ct." *RadioGraphics*, vol. 12, pp. 825–37, 1992.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980