Intracranial Hemorrhage Detection Using Transfer Learning

Abstract. The purpose of this project is to build a classification model that can accurately predict the probability of both the existence and subtype of an intracranial hemorrhage. We built two models, a binary classification model to detect the presence of a hemorrhage and a multi-class classification model to predict the subtype. We used transfer learning using the pretrained ResNet50 model with ImageNet input weights and were able to achieve an 89% test accuracy on the binary model and a 54% test accuracy on the multi-class model.

Keywords: intracranial hemorrhage · transfer learning · deep learning · head computed tomography

1 Introduction

Intracranial hemorrhage (ICH), or bleeding inside the brain, is a serious and sometimes deadly health problem that requires rapid identification and treatment [1]. Identifying the location and type of hemorrhage is a critical step before patient treatment can begin as consequences and treatment can vary extensively based on the size, type, and location of the hemorrhage. The location of the hemorrhage depicts its type: intraparenchymal, intraventricular, subarachnoid, subdural, and epidural. To detect and characterize these abnormalities, a radiologist uses a head Computed Tomography (CT) scan and determines the best course of treatment, which can sometimes mean immediate surgery. A CT scan produces three-dimensional (3D) stacks of images of the brain, where each cross section is one layer, or slice, of the brain. In each layer, brain tissues are captured with varying intensities, based on the amount of X-ray absorbency, and displayed with grayscale values. The grayscale values allow hemorrhage regions to appear denser and with a relatively undefined structure [2]. Once these images are captured, radiologists must sift through each one to detect an abnormality, note its location, and classify its type all in order to determine the best course of treatment. However, the detection of a hemorrhage is not only dependent on the presence of a radiologist who can analyze the scans, but it can also take a lot of time, which for many patients is already limited. Furthermore, if a radiologist misses even the smallest of abnormalities, that can prove fatal to the patient.

The existence of computer vision models that can detect information from medical images supports the idea that we can optimize this diagnosis process by creating a model to examine CT scans [3]. By using the main identifying features of a hemorrhage, such as density (hemorrhages appear whiter than the surrounding area), location, shape, and proximity to other cranial structures, the model will be able to predict the likelihood and subtype of the ICH [4]. Furthermore, abnormalities can occupy a very small amount of pixels, which can cause radiologists to overlook them [5]. However, since models are able to detect these minute abnormalities, they greatly minimize the time that radiologists need to spend analyzing the CT scans. Optimizing this diagnosis process will help the medical community treat patients more effectively, especially in areas where radiologists are not readily available. Through experimental results, our model has shown that transfer learning using ResNet50 as a pretrained model can be used in ICH diagnosis [6]; furthermore, our simple model is able to generalize with a large dataset and, with even more data, will be able to achieve better results.

2 Related Work

There has been much work on the problem of ICH detection with various approaches. The problem can be split into three subproblems, namely, detecting the existence of ICH, the subtype of ICH, and the location and volume of ICH (segmentation). The approaches to these subproblems vary in the type of learning, such as traditional, deep, or transfer learning, as well the size of the dataset.

Using traditional learning methods, Yuh et al. developed a threshold-based algorithm to detect ICH as well as ICH subtypes [7]. They were able to achieve 98% sensitivity and 59% specificity for the ICH detection while training and testing on a small dataset. In another work, Li et al. developed a method for the automatic detection of specifically subarachnoid hemorrhages [8]. Their method includes approximation of the subarachnoid space in a CT using an atlas-based registration. They were able to achieve a 100% sensitivity and 89.7% specificity on a small training and testing dataset.

Using a deep learning method, Kuo et al. trained a fully Convolutional Neural Network (CNN) they developed (PatchFCN) using 4,396 CT scans to attain expert-level detection of ICH [5]. Their algorithm demonstrated the highest accuracy to date for this clinical application, with a Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) of 0.99 ± 0.006 for identification of examinations positive for ICH. Their network is also able to perform segmentation, meaning it can identify the location and volume of ICH. Ye et al. developed a 3D joint convolutional and recurrent neural network (CNN-RNN) for the detection of ICH and its five subtypes [9]. Their network achieved greater than 0.98 AUC for the binary classification problem and greater than 0.8 AUC across all subtypes on a large dataset.

The transfer learning approach has been applied to other problems in the field of medical image analysis. Reddy et al. used transfer learning with ResNet50 for classification of malarial infected cells. Their experimental results show that transfer learning performs well on microscopic cell-images [10]. Tong et al. used transfer learning for ICH classification by employing three different types of CNN, LeNet[11], GoogLeNet[12], and Inception-ResNet [13]. They used a dataset consisting of 100 CT scans. Their model achieved accuracies of 0.997 for LeNet, 0.982 for GoogLeNet, and 0.992 for Inception-ResNet. Lee et al. used transfer learning by combining 4 CNN models, VGG-16[14], ResNet50[15], Inception-v3[17], and Inception-ResNet-v2, to detect ICH subtypes and locations [18]. Their model was trained on 904 CT scans and tested on a total of 437 CT scans, including both retrospective and prospective datasets. Their model achieved an AUC of 0.98 with 95% sensitivity and specificity for ICH detection. Their model achieved 78.3% sensitivity and 92.9% specificity for classification of the ICH subtypes.

3 Dataset

The dataset provided by the Radiological Society of North America (RSNA) contains images of CT scans in DICOM format[19]. DICOM, or Digital Imaging and Communications in Medicine, is a medical standard for communication of images and related data. Before we could begin the training phase, we needed to convert these images to PNGs or JPGs. We were able to utilize a pre-converted dataset that maintained the quality of the images.

The PNG dataset is split up into two parts: stage_1_train.csv file and the stage_1_train directory. The stage_1_train.csv file has two columns: the ID for the image in the stage_1_train directory and the type of hemorrhage. The stage_1_train directory contains all the images that we will be using to train our model. We ran a script on these files in order to create a directory for each hemorrhage type. We performed an 80:20 split on each hemorrhage directory to create separate train and test directories. Each directory contains 1600 train images and 400 test images. Before we could input this data into our model, we had to preprocess it. First, we performed a 90:10 split on our train dataset to create train and validation sets, leaving 1440 images for train and 160 images for validation. Each time we loaded the data, the validation and train sets were randomly chosen. Then we loaded images as arrays of size 224×224 pixels because the ResNet50 model is trained on images of this size.

For the binary classification model we encoded the labels as 0 ('no') or 1 ('yes'). Then we created generators for each set - train, validation, and test - but for the training set we used Image Data Generator to perform transformations on each image, such as stretching, zooming, and more. This creates multiple batches of tensor image data, not only allowing us to increase our train set, but also to apply real-time data augmentation. For the multi-class classification model we encoded each of the five labels as numeric values from 0 to 4 and then converted those to one hot-encoded labels. This ensured that there was no bias towards higher values. Then we created the same type of generators as the binary model and performed the same data augmentation on the train set.

4 Models and Algorithms

To accurately identify ICH, our models needed to extract the main features of a hemorrhage: density, location, shape and proximity. There are various existing

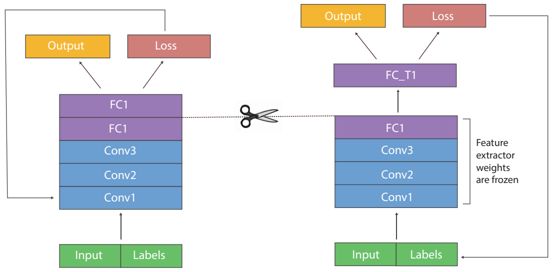

Fig. 1. A visualization of our transfer learning technique. FC1 = fully connected layer 1, FC_T1 = fully connected transfer layer 1 (our added fully connected layer).

pretrained deep neural networks for such computer vision problems, such as VGG-19, Inception V3, Xception, and ResNet50 [15]. These networks are able to generalize images outside the ImageNet[16] dataset via transfer learning, through methods such as feature extraction and fine-tuning. By using these input weights, we were able to extract features to use on our dataset, which decreased our model's training time. Additionally, we determined, in order to fully encapsulate the scope of the classification problem, we needed to implement two models: binary classification and multi-class classification.

Both models begin with the same structure: transfer learning using the pre-trained ResNet50 model with ImageNet input weights. A visualization of our transfer learning technique can be seen in Figure 1. We used transfer learning because we did not have enough data to extract features from our images, so using this allowed us to transfer weights from a preexisting data pool to our model. To effectively apply this, we needed to transfer the basic features of an image such as shape and illumination. These are particularly helpful for our problem because these are the two main features we look at when identifying hemorrhages. For the binary classification model, the illumination feature was particularly important because hemorrhages appear lighter than the rest of the cranial structures in the CT scan. For the multi-lass model, since we already knew a hemorrhage existed in the scan, the location and shape features were more important for the model to extract.

When we loaded the ResNet50 model, we froze all of its layers except the Batch Normalization layers. By making these layers trainable, we allow weight updates to happen during training. We also do not load the ResNet50 model's final fully connected layer because we add our own so that we can update the parameters of our dataset. Before we added this final layer, we added a Global Average Pooling layer and our final fully connected layer: a dense layer with 256 nodes and a Relu activation function. Then, as a regularization method, we use a dropout layer with a fraction rate of 0.4 and, after, add a Batch Normalization layer.

Then, we added the final output layer, which is where the differences between our two models begin. For the binary model, our output layer consisted of a dense layer with one node and a sigmoid activation function since we are classifying the images as either containing a hemorrhage or not. When we predicted the class of an image, this model outputted a value between 0 and 1, which can directly be translated to the probability of the CT scan containing a hemorrhage. Our metrics to evaluate this model were accuracy and binary cross-entropy loss. For the multi-class model, our output layer consisted of a dense layer with five nodes and a softmax activation function since we had multiple classes. When we predict on this model, it outputs five probabilities, one for each hemorrhage sub-type. Our metrics for this model were accuracy and categorical cross-entropy loss.

Initially, we trained our models with a dropout rate of 0.25, a training batch size of 8, a Relu non-linearity function, and the Adam optimization method. To ensure that these values provided us with the best accuracy, we optimized these hyper parameters by training both models repeatedly with various hyper parameters and analyzing which combination gave us the best output. For the binary model, we ended up modifying just the dropout rate to 0.4 and the training batch size to 16. For the multi-class model, we only modified the dropout rate to 0.3. By conducting these hyper parameter optimizations, we were able to make sure that we were training our models using the best values possible.

5 Results and Analysis

5.1 Binary Classification Model

To quantitatively evaluate our model, we used binary cross-entropy log loss which measures the performance of a classification model with only two outputs; as the predicted probability diverges from the actual label, the loss increases. As seen in Figure 3, the training phase of the model is almost convergent as the loss decreases over epochs and, as seen in Figure 2, the accuracy converges as the training period continues. Thus, our model is classifying the existence of a hemorrhage more accurately as it continues to train, and, after the training period, is able to accurately classify ICH 89% of the time, as seen in Table 1.

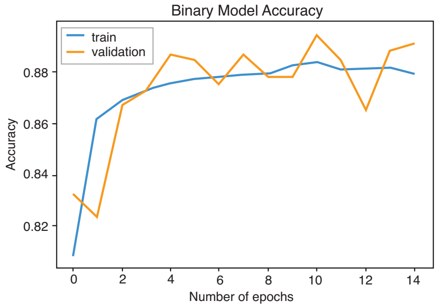

Fig. 2. Accuracy versus Number of Epochs of the Binary Classification Model

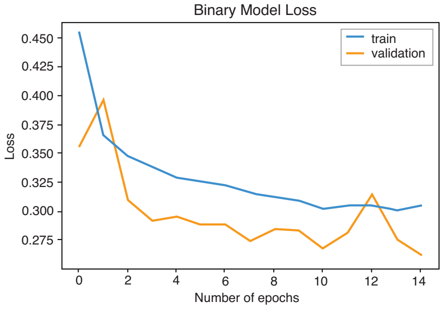

Fig. 3. Loss versus Number of Epochs of the Binary Classification Model

Table 1. Loss and Accuracy of Train, Validation, and TestSets for the Binary Classification Model

	Loss		**Accuracy**	
	Start	**End**	**Start**	**End**
Train	0.45	0.30	0.81	0.88
Validation	0.35	0.26	0.83	0.89
Test	0.27		0.89	

Another evaluation metric we utilized was a confusion matrix to identify the amount of correctly classified inputs. In Table 2, we can see that the majority of 'yes', or 1, labels are predicted accurately compared to 'no' labels. While this may be due to a bias in our dataset, where we have significantly larger numbers of 'yes' data than 'no', it still supports the idea that our model is able to classify ICH. This model also achieved a testing sensitivity of 92% and specificity of 72%. With such a high sensitivity, we can see that our model rarely misdiagnoses the presence of a hemorrhage.

Table 2. Confusion Matrix for the Binary Classification Model

		Predicted Labels	
		0	**1**
Actual	**0**	227	173
Labels	**1**	87	1913

5.2 Multi-class Classification Model

To quantitatively evaluate this model, we used a categorical cross-entropy log loss, since we had more than two exclusive classes, a softmax activation function, and used one-hot encoded labels. Our model is able to learn as the accuracy improves over epochs, as we can see in Figure 4, and the loss of our model trends downwards, as we can see in Figure 5. As seen in Table 3, our model achieves a testing accuracy of 53%, which although low, is reasonable since the differences between the subtypes of ICH are minute and can be hard to distinguish, even for a human.

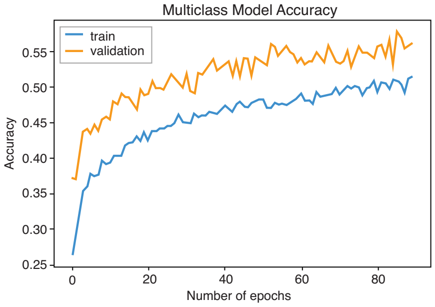

Fig. 4. Accuracy versus Number of Epochs of the Multi-class Classification Model.

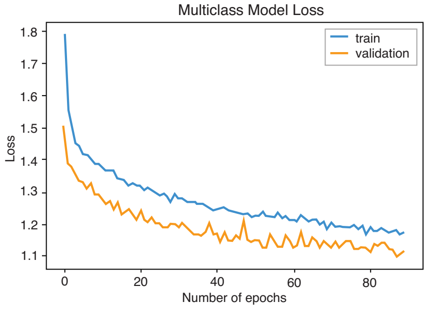

Fig. 5. Loss versus Number of Epochs of the Multi-class Classification Model

We also used a confusion matrix, Table 4, to evaluate this model and can see that the model is able to classify many inputs accurately. The hemorrhage subtype with the best performing testing sensitivity was intraventricular, with a value of 62%; however, there was not much of a difference between the sensitivities of the rest of the subtypes, which as mentioned earlier, stems from the minute differences between subtypes. The epidural class had a sensitivity of 53%, the intraparenchymal class had 57%, the subarachnoid class had 48%, and the subdural class had 50%.

6 Future Work

The main improvement we would implement would be to create the multilabel multi-class classification model. With this, we would be able to predict at once

Table 3. Loss and Accuracy of Train, Validation, and Test Sets for the Multi-class Classification Model

	Loss		**Accuracy**	
	Start	**End**	**Start**	**End**
Train	1.79	1.17	0.26	0.51
Validation	1.50	1.11	0.37	0.56
Test	1.14		0.53	

Table 4. Confusion Matrix for the Multi-class Classification Model

		Predicted Labels				
		0	**1**	**2**	**3**	**4**
Actual Labels	**0**	225	32	9	54	80
	1	43	189	86	54	28
	2	25	68	235	47	25
	3	47	29	30	219	75
	4	84	13	18	80	205

0 = epidural, 1 = intraparenchymal, 2 = intraventricular, 3 = subarachnoid, 4 = subdural

if a model contains a hemorrhage and, if so, of which type. Currently, to make those predictions, a data point should be presented to both models, which is less efficient.In addition, while we chose ResNet50 as our pretrained model due to various reasons mentioned earlier, we would also like to test other pretrained models so that we can properly compare the results between each one.

To further test the accuracy of our model, we would have also liked to compare the results of our model to that of a radiologist. By doing so, we can see how our model performs compared to a radiologist and if it actually makes the process more efficient. With the lack of access to radiologists and a large dataset that would make the process time consuming, we were unable to make this comparison, but would like to do in the future. Finally, an additional feature that would be useful for doctors would be to place a border box around the hemorrhage, clearly highlighting its location and size. Since our model already outputs the likelihood of the subtype of hemorrhage, these are the only additional features a doctor would need in order to make a diagnosis.

7 Conclusion

We built a binary and a multi-class classification model to detect and classify ICH. We used transfer learning using the pretrained ResNet50 model with ImageNet input weights and were able to achieve an 89% test accuracy on the binary model and a 54% test accuracy on the multi-class model. Even though our models do not have perfect test accuracy, they still present potential for diagnosing

ICH. However, even if perfect accuracy is achieved, the next hurdle is to seamlessly integrate this technology into the everyday workflow of a radiologist. In conclusion, this technology could optimize the process and decrease the amount of time between taking a CT scan and beginning treatment for a hemorrhage, especially in places where radiologists are a limited resource.

References

1. Ojemann, R.G., Heros, R.C.: Spontaneous Brain Hemorrhage. Stroke, vol. 14, pp. 468–475. (1983). https://doi.org/10.1161/01.STR.14.4.468
2. Hssayeni, M. D., Croock, M. S., Salman, A. D., Al-khafaji, H. F., Yahya, Z. A., Ghoraani, B.: Intracranial Hemorrhage Segmentation Using A Deep Convolutional Model. Data, vol. 5(1), pp. 1–18. (2020). https://doi.org/10.13026/w8q8-ky94
3. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I. A survey on deep learning in medical image analysis. Medical image analysis, vol. 42, pp. 60–88. (2017). https://doi.org/10.1016/j.media.2017.07.005
4. RNSA Intracranial Hemorrhage Detection, `https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection`. Last accessed 12 March 2020
5. Kuo, W., Hne, C., Mukherjee, P., Malik, J., Yuh, E. L.: Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proceedings of the National Academy of Sciences, vol. 116(45), pp. 22737–22745. (2019). https://doi.org/10.1073/pnas.1908021116
6. Intracranial Hemorrhage Detection Using Transfer Learning, `https://github.com/ivpetkov/ich-detection`. Last accessed 2 April 2020
7. Yuh, E.L., Gean, A.D., Manley, G.T., Callen, A.L., Wintermark, M.: Computer-aided assessment of head computed tomography (CT) studies in patients with suspected traumatic brain injury. Journal of neurotrauma, vol. 25(10), pp. 1163–1172. (2008). https://doi.org/10.1089/neu.2008.0590
8. Li, Y., Wu, J., Li, H., Li, D., Du, X., Chen, Z., Jia, F., Hu, Q.: Automatic detection of the existence of subarachnoid hemorrhage from clinical CT images. Journal of medical systems, vol. 36(3), pp. 1259–1270. (2012). https://doi.org/10.1007/s10916-010-9587-8
9. Ye, H., Gao, F., Yin, Y., Guo, D., Zhao, P., Lu, Y., Wang, X., Bai, J., Cao, K., Song, Q., Zhang, H.: Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. European radiology, vol. 29(11), pp. 6191-6201. (2019). https://doi.org/10.1007/s00330-019-06163-2
10. Reddy, A.S.B., Juliet, D.S.: Transfer Learning with ResNet-50 for Malaria Cell-Image Classification. 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0945–0949. IEEE. (2019). https://doi.org/10.1109/ICCSP.2019.8697909
11. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. IEEE, vol. 86(11), pp. 2278–2324, (1998).
12. Szegedy, C. et al.: Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, (2015).
13. Phong, T.D., Duong, H.N., Nguyen, H.T., Trong, N.T., Nguyen, V.H., Van Hoa, T., Snasel, V.: Brain hemorrhage diagnosis by using deep learning. In Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, pp. 34-39. (2017). https://doi.org/10.1145/3036290.3036326

14. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556 (2014).
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 7, pp. 770–778. (2015) https://doi.org/10.1109/CVPR.2016.90
16. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. (2009)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 (2015).
18. Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S.H., Guerrier, C.E., Ebert, S.A., Pomerantz, S.R., Romero, J.M., Kamalian, S., Gonzalez, R. G.: An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nature Biomedical Engineering, vol. 3(3), pp. 173. (2019). https://doi.org/10.1038/s41551-018-0324-9
19. RNSA Intracranial Hemorrhage Detection Detection Stage 1 PNG 128x128x Dataset, https://www.kaggle.com/guiferviz/rsna_stage1_png_128. Last accessed 4 February 2020
20. Chollet, F.: Keras. https://keras.io (2015)
21. Kulkarni, B: Transfer Learning with ResNet. https://medium.com/@balaji.kulkarni92/transfer-learning-using-resnet-e20598314427 (2019)
22. Khandelwal, R.: Deep Learning using Transfer Learning. https://towardsdatascience.com/deep-learning-using-transfer-learning-python-code-for-resnet50-8acdfb3a2d38 (2019)