MATH7243 Final Project - Brain Hemmorhage Image Classification/Segmentation

Broderick Kelly, Mitchell Whelan, Minoo Mohebbifar

April 23, 2025

Abstract

Intracranial hemorrhage (ICH) is a critical medical condition requiring swift and accurate detection for effective patient management. This study explores a range of classical machine learning and advanced deep learning approaches to classify and segment brain hemorrhages from CT scan images. Initially, classical models such as Logistic Regression, k-Nearest Neighbors (kNN), Decision Trees, Random Forests, Quadratic Discriminant Analysis (QDA), and Support Vector Machines (SVM) with an RBF kernel were implemented. While classical methods demonstrated limited effectiveness, with maximum accuracy reaching only 50% (QDA), deep learning techniques showed significantly improved performance. Convolutional Neural Networks (CNNs), particularly transfer-learning models based on ResNet50 and Xception architectures, substantially enhanced classification accuracy, achieving up to 80.16% accuracy for binary hemorrhage detection and 65.86% for seven-class hemorrhage subtype categorization. Segmentation tasks were addressed using a ResNet50-based U-Net architecture, which achieved high precision (77.32%) and recall (71.65%) in delineating hemorrhage regions. The results highlight the clear superiority of deep learning models for both classification and segmentation tasks in brain CT imaging, underscoring their potential to assist clinicians in emergency neuroradiology settings.

Introduction

Intracranial hemorrhage (ICH) is a life-threatening emergency caused by bleeding within the skull, often due to trauma or stroke. Rapid identification and characterization of brain hemorrhages on non-contrast CT scans are critical for timely intervention [3]. Radiologists traditionally diagnose and classify hemorrhages by examining CT images, identifying subtypes such as epidural (EDH), subdural (SDH), subarachnoid (SAH), intraparenchymal (also known as intracerebral, ICH or IPH), intraventricular (IVH), or cases with multiple concurrent hemorrhages, as well as distinguishing normal (no hemorrhage) scans.

Correctly classifying the hemorrhage subtype is important because each type can require a different management strategy (e.g., EDH often demands urgent surgical evacuation, whereas SAH may require aneurysm securing). Moreover, quantifying the extent of bleeding (e.g., volume of a hematoma) via segmentation is vital for prognosis and surgical planning [4, 5].

However, manual analysis of CT scans is labor-intensive and subject to inter-observer variability, especially under emergency conditions or in resource-limited settings where experienced neuroradiologists may not be readily available [3]. This has motivated the development of computer-aided diagnosis tools to assist in hemorrhage detection, classification, and segmentation.

Early attempts at automated hemorrhage identification used classical machine learning (ML) approaches with hand-crafted features. Techniques such as support vector machines (SVMs), decision trees, random forests, logistic re-

gression, and other classifiers have been applied to detect hemorrhagic lesions on head CT [6, 7]. In these approaches, radiomic features (e.g., intensity histograms, texture measures, shape descriptors) are extracted from the CT images and fed into ML models to differentiate hemorrhage types. For instance, a radiomics-based SVM model on a limited dataset could achieve around 85–90% accuracy in distinguishing hemorrhage subtypes [6].

While such classical methods demonstrated the potential of automation, they were often constrained by the need for meticulous feature engineering and suffered from the "curse of dimensionality" when the feature set was large relative to the number of training samples [6]. In practice, deep learning approaches have largely surpassed classical methods in performance on this task [3]. A recent meta-analysis concluded that modern deep learning models can detect ICH on par with expert radiologists, with pooled sensitivity around 0.89, specificity 0.91, and area under the ROC curve around 0.94 [3]. These models can also segment hemorrhagic lesions with high accuracy (average Dice coefficient $\approx$ 0.84), enabling automated volume measurements that agree closely with manual annotations [3].

Such results underscore the promise of advanced machine learning for rapid and reliable hemorrhage assessment.

Project Objective: In this project, we leverage both classical and deep learning techniques to perform comprehensive hemorrhage analysis on brain CT images. The goal is to develop models that can classify CT slices into one of seven categories – epidural, subdural, subarachnoid, intraparenchymal, intraventricular, multiple hemorrhage, or normal – and to segment the hemorrhagic region for positive cases. This extends prior works (such as the RSNA 2019 challenge, which defined five subtypes of intracranial hemorrhage) by explicitly handling cases of mixed hemorrhage (the "multi" class) and providing pixel-level delineation of bleeding [8].

We explore a spectrum of machine learning approaches: from baseline methods like logistic regression, k-nearest neighbors (k-NN), quadratic discriminant analysis (QDA), and SVMs with kernel tricks, to state-of-the-art deep neural networks. In particular, we investigate convolutional neural network (CNN) architectures (including a custom CNN, ResNet-50, and Xception backbones) for the multi-class classification task, as well as a U-Net–based CNN for segmentation of hemorrhagic lesions. By comparing these approaches, we aim to highlight the trade-offs between classical ML models (which are simpler and interpretable but may be less accurate) and advanced deep learning models (which typically achieve higher accuracy at the cost of requiring more data and computation).

Ultimately, our work seeks to demonstrate an effective pipeline for brain CT hemorrhage classification and segmentation, contributing to the development of AI tools that could assist clinicians in emergency neuroradiology settings.

Related work

Research on brain CT hemorrhage has progressed from early classical machine learning models to advanced deep learning systems.

Classical ML and Hybrid Methods: Traditional models using handcrafted radiomic features and classifiers like logistic regression, SVMs, and random forests provided moderate performance in multi-class hemorrhage classification. Hybrid approaches combining radiomic and deep features improved accuracy, but classical models often struggle with generalization.

Deep Learning for Classification: Deep CNNs, particularly pretrained networks like ResNet and DenseNet, have become standard in hemorrhage classification, achieving high accuracy on benchmark datasets. Recent models, including Vision Transformers, have demonstrated even higher performance by capturing complex spatial patterns.

Deep Learning for Segmentation: U-Net and its variants remain dominant in hemorrhage segmentation, achieving high Dice scores and reliable volume estimation. Enhanced versions like attention-based U-Nets further improve localization accuracy. 3D U-Nets also show promise for complex and small-volume bleeds.

Summary: Overall, deep learning methods outperform classical approaches in both detection and segmentation. Our project builds on these developments by comparing classical and deep models for classifying and segmenting CT-based hemorrhages.

Method description

Application of Logistic Regression for Classification

Logistic Regression was implemented as a baseline discriminative classifier to determine whether a CT scan image reveals intracranial hemorrhage. The binary target variable was again the *"any"* column, identifying the presence or absence of bleeding.

To deal with the multi-classification, we implemented 6 different binary classification models and combined them to get an overall model. The individual class accuracy was generally pretty good, but the overall accuracy was low.

Logistic Regression obtained an overall accuracy of 15%. However, this method still struggled to capture the complex, nonlinear patterns present in the image data, suggesting the need for more flexible classifiers. The confusion matrix for each binary classification model is below:

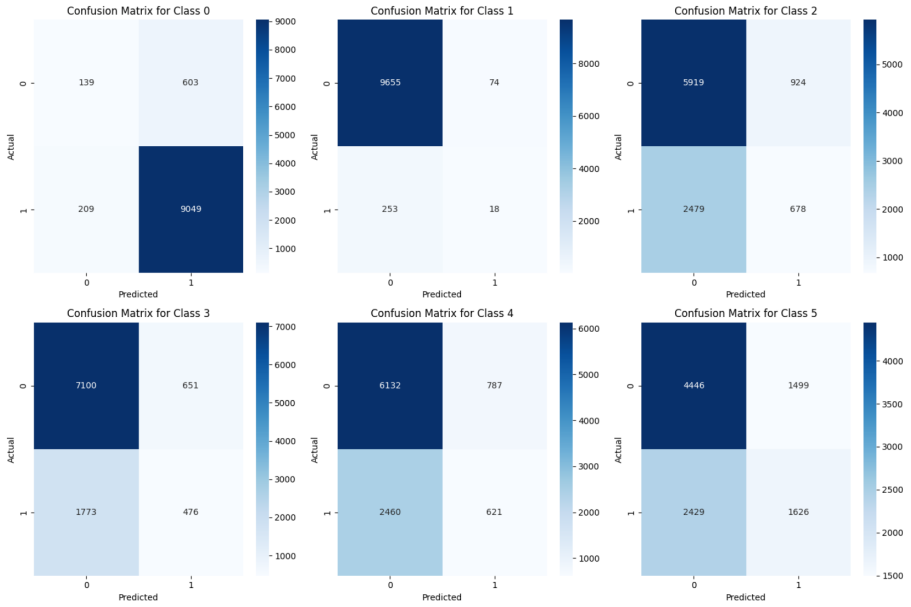

Figure 1: Confusion Matrix for Logistic Regression

To contextualize the limitations of this model, we outline the mathematical basis of Logistic Regression below.

Logistic Regression – Theoretical Overview

Logistic Regression is a discriminative model that estimates the posterior probability $P(y = 1 \mid x)$ directly using the logistic (sigmoid) function:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-x^\top \beta)}$$

The log-odds (logit) function is modeled as a linear combination of input features:

$$\log\left(\frac{P(y = 1 \mid x)}{1 - P(y = 1 \mid x)}\right) = x^\top \beta$$

Model parameters β are estimated by maximizing the likelihood function over the training data. For binary classification, the predicted class label is:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1 \mid x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Application of k-Nearest Neighbors (kNN) for Classification

k-Nearest Neighbors (kNN) was explored as a non-parametric method for classifying CT scan images based on similarity to labeled examples. Each test image was assigned the majority label among its k nearest neighbors in the training set, using Euclidean distance in the feature space.

Due to the high dimensionality of image data, we applied dimensionality reduction (e.g., PCA) before running kNN to reduce noise and improve computational efficiency.

While kNN captured some local structure in the data, it was sensitive to the choice of k and the curse of dimensionality. The best-performing configuration yielded an accuracy of 25% with very good individual binary classifier accuracy. The confusion matrix is provided below:

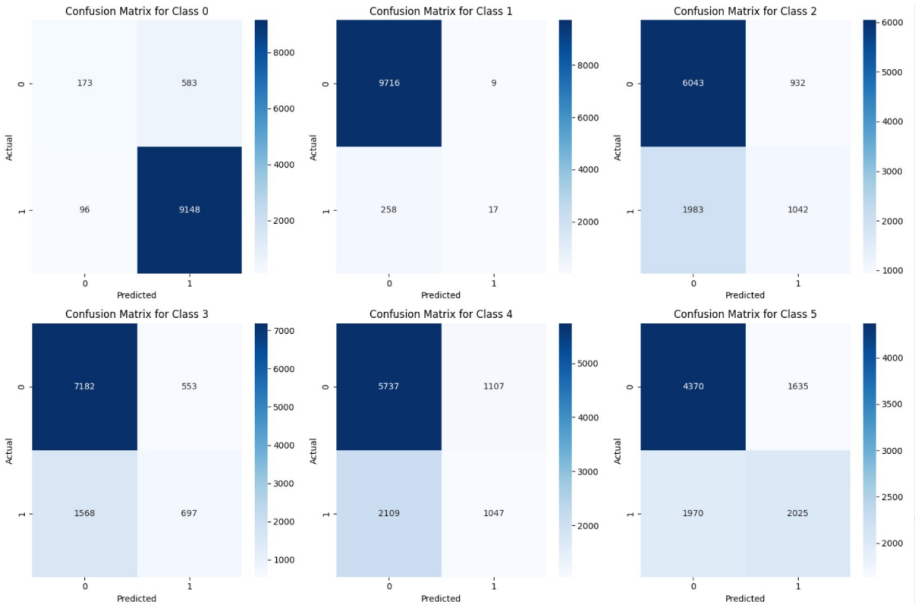

Figure 2: Confusion Matrix for kNN Classification

A brief overview of the theoretical foundation of kNN is provided below.

k-Nearest Neighbors (kNN) – Theoretical Overview

The kNN classifier assigns a class label to a new observation x by majority vote among the k nearest training points:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i \in \mathcal{N}_k(x)} I(y_i = c)$$

where $\mathcal{N}_k(x)$ denotes the indices of the k closest points to x and $I(\cdot)$ is the indicator function. Distance is typically measured using Euclidean distance:

$$\text{dist}(x, x_i) = \|x - x_i\|_2$$

kNN is a memory-based method and does not involve explicit training. Its performance heavily depends on feature scaling and dimensionality.

Application of Tree-Based Methods for Classification

To address the nonlinear nature of the CT scan image data, we applied Decision Trees and their ensemble extensions such as Random Forests. These methods partition the feature space into regions associated with specific class labels, capturing complex interactions and nonlinearities.

Training was conducted on a balanced dataset using the "subdural_window" images. We tuned hyperparameters such as tree depth and number of estimators (for Random Forests) using cross-validation.

Tree-based methods significantly outperformed earlier models, with Decision Trees obtaining an accuracy of 20% and Random Forests achieving an accuracy of 11.5%. Their ability to model complex patterns and inherent resistance to outliers made them highly suitable for this task. The confusion matrix for the Decision Tree model is below:

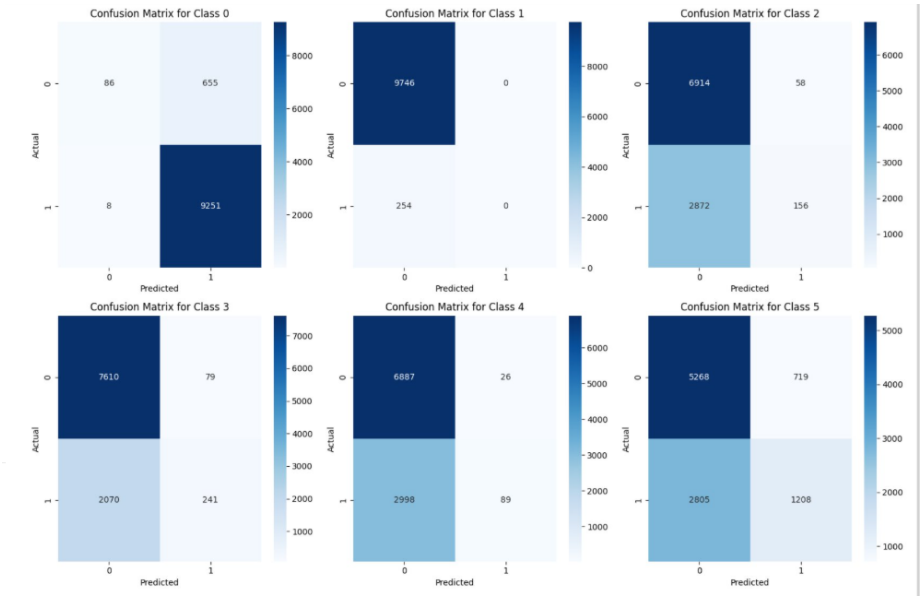

Figure 3: Confusion Matrix for the Decision Tree Model

The confusion matrix for the Random Forest model is below:

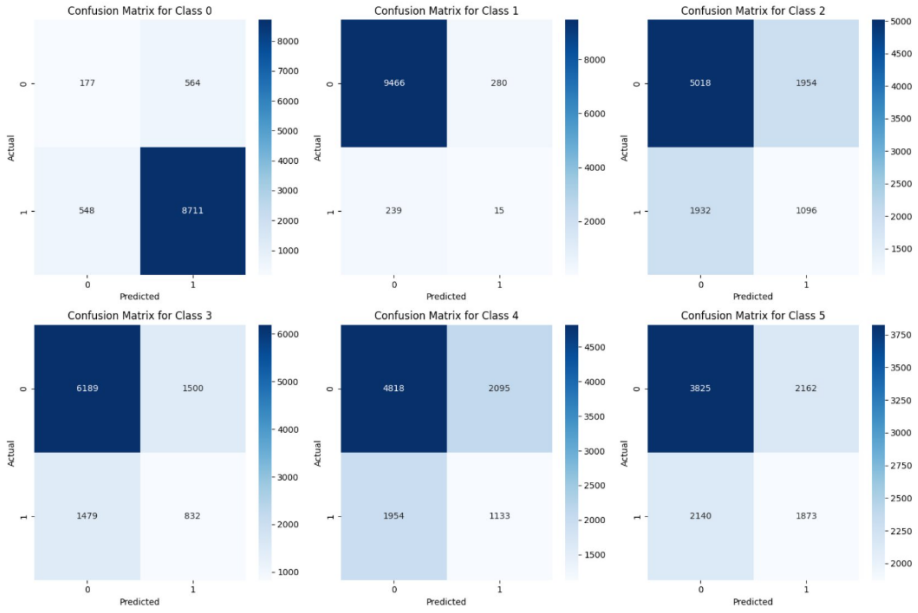

Figure 4: Confusion Matrix for the Random Forest Model

We summarize the underlying theory of Decision Trees below.

Tree-Based Methods – Theoretical Overview

Decision Trees recursively split the feature space to minimize a loss function such as the Gini impurity or entropy. For a node with class proportions $p_1, p_2, \ldots, p_K$, the Gini impurity is:

$$G = \sum_{k=1}^{K} p_k(1 - p_k)$$

The optimal split at each node is chosen to maximize the reduction in impurity:

$$\Delta G = G_{\text{parent}} - \left(\frac{n_L}{n} G_L + \frac{n_R}{n} G_R\right)$$

For classification, the predicted label is the majority class of the leaf node into which a sample falls:

$$\hat{y} = \arg\max_k \text{ proportion of class } k \text{ in leaf}$$

Random Forests improve generalization by aggregating predictions from multiple decorrelated trees trained on bootstrapped data and random feature subsets.

Application of Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) was employed as an initial method to classify whether a CT scan image exhibits bleeding. For this task, we used the *"any"* column of the data set as the target variable, which indicates the presence or absence of hemorrhage.

One major challenge encountered was the discrepancy in dataset sizes. The label data set contained annotations for 750,000 images, whereas the available image dataset included only 463,000 images, primarily from cases of bleeding. This imbalance introduced a bias in the model, leading it to favor the bleeding class. To address this, we balanced the training dataset by ensuring an equal number of images from both classes.

Each sample in the dataset included four different image types. Initially, we trained four separate QDA models, each using a different image type. However, these models did not converge, which led us to simplify our approach by selecting only the "subdural_window" image type for classification. Even within this subset, a severe class imbalance was observed—only 8,000 normal cases compared to 108,000 bleeding cases. To mitigate this issue, we randomly selected 8,000 images from the bleeding cases to match the 8,000 normal cases, forming a balanced training set.

After training the QDA model on this balanced dataset, the classifier achieved an accuracy of 50%, indicating that QDA was ineffective for this complex image classification task. The low performance suggests that QDA's assumptions are not well-suited for high-dimensional image data, motivating the need for more advanced classification techniques.

To better understand the limitations observed in our experiments, we provide a brief overview of the mathematical formulation of Quadratic Discriminant Analysis (QDA) below.

Quadratic Discriminant Analysis (QDA) – Theoretical Overview

Quadratic Discriminant Analysis (QDA) is a generative classification method based on Bayes' theorem. It models each class conditional distribution as a multivariate Gaussian with its own covariance matrix. The class-conditional density is given by:

$$p(x \mid y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)\right)$$

where μ_k and Σ_k denote the mean and covariance matrix of class k, respectively. Using Bayes' theorem, the posterior probability becomes:

$$P(y = k \mid x) = \frac{p(x \mid y = k)P(y = k)}{p(x)}$$

In practice, the following discriminant function is used for classification:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

where π_k is the prior probability of class k. The predicted label is:

$$\hat{y} = \arg\max_k \delta_k(x)$$

Support Vector Machine (SVM) with RBF Kernel

Support Vector Machines (SVMs) are supervised learning models used for binary classification by finding the optimal separating hyperplane between two classes. In our project, we used SVM with a Radial Basis Function (RBF) kernel to classify subdural-window brain CT images into hemorrhage and non-hemorrhage cases.

The standard form of the SVM classifier solves the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where:

- $\mathbf{x}_i$ are the training samples, $y_i \in \{-1, 1\}$ are the labels,
- $\phi(\cdot)$ maps inputs to a higher-dimensional feature space,
- $C > 0$ is the regularization parameter,
- ξ_i are slack variables that allow for misclassification.

We used the RBF kernel, defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

The RBF kernel allows the decision boundary to be nonlinear in the input space, making it suitable for the high-dimensional patterns in medical images. The γ parameter controls the spread of the kernel; small γ results in smoother decision boundaries.

In our implementation, we used a simplified version of the QDA setup: a binary classification task on balanced data, selecting 8,278 hemorrhage and 8,278 non-hemorrhage images. While QDA showed poor performance (50% accuracy), the SVM model trained on the same data with the RBF kernel improved generalization and decision boundary flexibility. The kernel SVM allowed us to capture nonlinear patterns in the subdural-window CT scans, making it more suitable for this classification task.

However, when extended to the full 7-class hemorrhage classification problem, the SVM model with RBF kernel failed to converge and did not generalize well. The model was not able to handle the complexity and class imbalance inherent in the multi-class dataset, indicating a limitation of SVMs in this high-dimensional, multi-label medical imaging scenario.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are particularly well-suited for hemorrhage classification in medical imaging tasks. These networks leverage convolutional layers to automatically extract spatial and intensity-based features from the images—such as the shape, location, and density variations characteristic of different types of hemorrhages (e.g., intraparenchymal, subdural, or subarachnoid). CNNs excel at capturing local dependencies in the image through small, learnable filters, making them ideal for identifying subtle patterns in pixel intensity that may indicate the presence of blood, edema, or tissue displacement.

One of the key advantages of CNNs in this context is their ability to learn robust, hierarchical feature representations without manual intervention. Early layers may detect simple edges or texture gradients, while deeper layers capture complex structures like hematoma boundaries or midline shifts. This hierarchical representation is crucial in medical imaging, where clinical significance often depends on both fine-grained features and their broader anatomical context. Additionally, existing CNN architectures can be adapted for classification, localization, or even segmentation tasks, allowing flexibility in designing a pipeline for automated hemorrhage detection.

In this study CNNs were applied in two ways, for image classification and for image segmentation. Prior models in this study were exclusively focused on image classification, attempting to categorize the hemorrhaging into one of seven types. Image segmentation refers to generating binary masks that delineate the regions affected by hemorrhaging. For the CNN applications, data representing the brain, bone and subdural windows were combined as RGB images to provide multi-channel input that enhances contrast. These composite images were used to train and evaluate various CNN architectures to identify those most effective for hemorrhage detection and classification.

Application of CNNs for Classification

For all classification approaches, the input images were down-sampled from 512-by-512 to 256-by-256, to reduce computation time and memory usage without losing significant structural detail. Future studies could experiment with different down-sampling ratios to achieve the best balance of accuracy and efficiency. All modeling was done with an equal number of images in each class, which was about 2,000 images for each class in the 7-class models and 7,000 for each class in the 2-class models. As a baseline, a ten-layer basic CNN was built and trained, and achieved a test accuracy of 15.02%, barely more accurate than random choice between the seven classes (14.29%). Due to the complexity of

the task and time and processing constraints, the rest of the implementations utilized transfer learning, using pre-trained weights from ResNet50 and Xception. ResNet50 is a powerful, deep CNN architecture built and introduced by Microsoft Research in 2015. It consists of 50 layers and utilizes residual blocks, enabling the network to skip connections and avoid vanishing gradient problems [1]. Xception, is another deep CNN known for novel feature learning, released by Google in 2017. It leverages depth-separable convolutions, which reduces the number of its parameters and allows it to consist of only 36 convolutional layers while maintaining great performance [2]. Both models are extremely capable, achieving 92.2% and 95.5% respectively, in top-5 classification accuracy on the ImageNet benchmark dataset. These pre-trained models were used as the backbone for models in this study, based on a modified U-Net architecture adapted for classification, with 5 additional layers added to transform these general models to specific hemorrhage classification models.

For ResNet50, skip connections were extracted from intermediate layers conv1_relu, conv2_block3_out, conv3_block4_out, conv4_block6_out with the deepest feature map taken from conv5_block3_out. For Xception, features were taken from block2_sepconv2_bn, block3_sepconv2_bn, block4_sepconv2_bn, and block13_sepconv2_bn, with the bridge feature map from block14_sepconv2_bn. These skip connections were retained for later concatenation during up-sampling. The decoding upsampling blocks were implemented using Conv2DTranspose layers, each followed by concatenation with the corresponding encoder skip connection, convolution using 3x3 filters, batch normalization, ReLU activation and dropout. Then after the decoder there is a classification head that collapses the spatial dimensions into a 1D feature vector which is passed through a Dense(256) layer with ReLU and 40% dropout, and a Dense(128) layer with ReLU and 30% dropout. The final layer is a dense layer with softmax activation producing class probability outputs. The models were compiled with an Adam optimizer with a learning rate of 0.001, and the data were randomly split into training, validation, and testing data with a split of 70/20/10. The model was trained with the pre-trained weights frozen for 20 epochs, and then an additional 5 epochs with the pre-trained weights unfrozen for fine adjustments.

The CNN built with a ResNet50 backbone achieved test accuracy of 30.29%, about double the baseline accuracy. The results are visualized in the confusion matrix below.

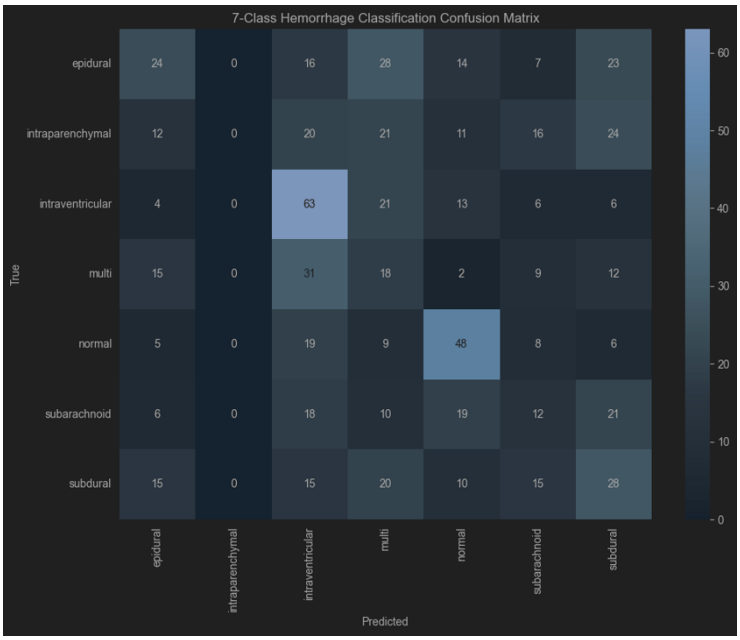

Figure 5: Confusion Matrix of ResNet50 7-class categorization

Some classes perform much better than others, with intraventricular and normal classes being more easily identified by the model. Additionally, there is a dead neuron effect for the intraparenchymal class, where despite the classes being balanced, the model never predicted the intraparenchymal class as the correct class. This means that during the final softmax function, the probability of the intraparenchymal class was never greater than the other classes. This could be due to low feature separability or vanishing gradients during training. Due to surprisingly poor performance and strange class disparities, the multi class, which represented scans with multiple hemorrhage types, was removed in an attempt to reduce label ambiguity. However, this adjustment did not produce improved performance, and the new model achieved a functionally equivalent test accuracy of 31.40%. The confusion matrix for this model is shown below.

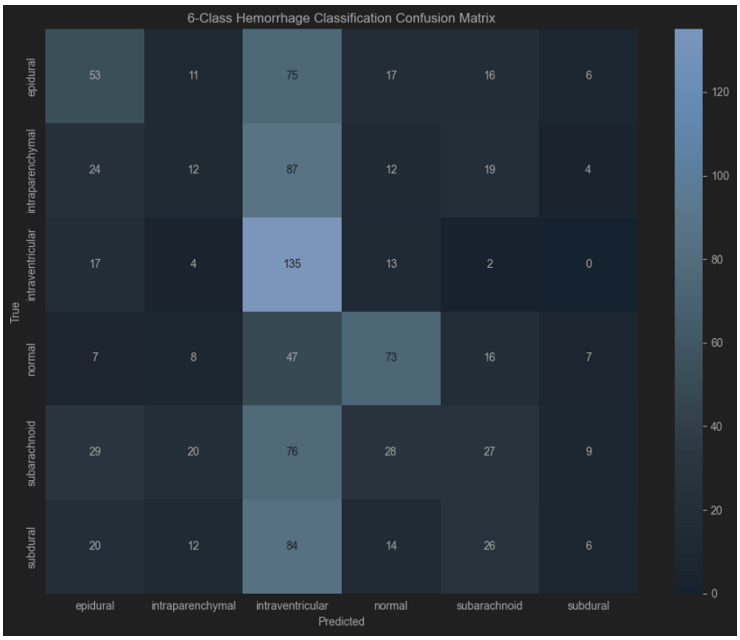

Figure 6: Confusion Matrix of ResNet50 6-class categorization

Additionally, in order to simplify the problem, a model was built to simply identify if an image had hemorrhaging in it by identifying if an image was normal or hemorrhaged. To do this, about 1,000 images from each hemorrhage class were combined into a master hemorrhage class. 7,000 normal images were used as well to achieve balanced classes, and then the model was run. Using the ResNet50 backbone, the model achieved an accuracy of 70.54%. Although performance improved significantly, this is largely due to the simplification of the classification task from seven categories to two, and there is still room for improvement. The confusion matrix for this model is shown below.

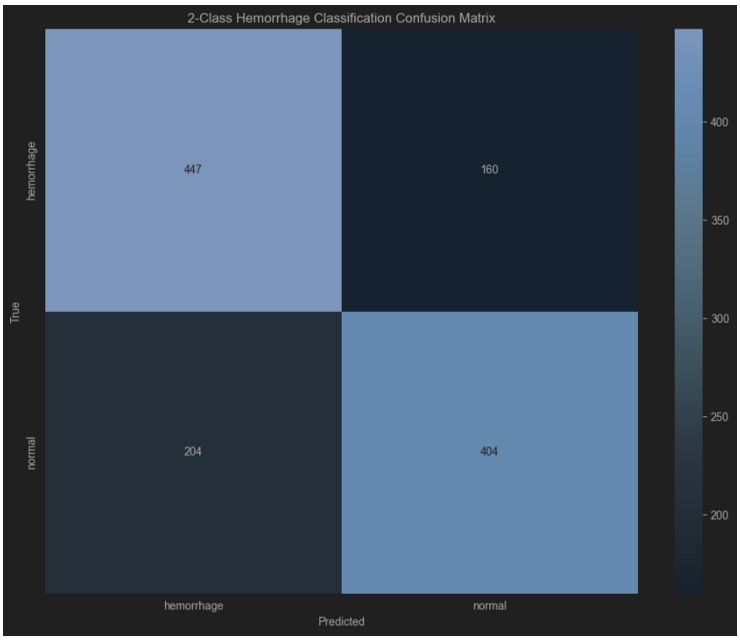

Figure 7: Confusion Matrix of ResNet50 binary categorization

CNNs built with an Xception backbone achieved much better test accuracy than their ResNet50 counterparts. For the 7-class categorization problem, the Xception model achieved a test accuracy of 65.86%. A binary normal/hemorrhaged model was also built with Xception and this outperformed the ResNet50 binary model, achieving a test accuracy of 80.16%. Both confusion matrices are shown below.

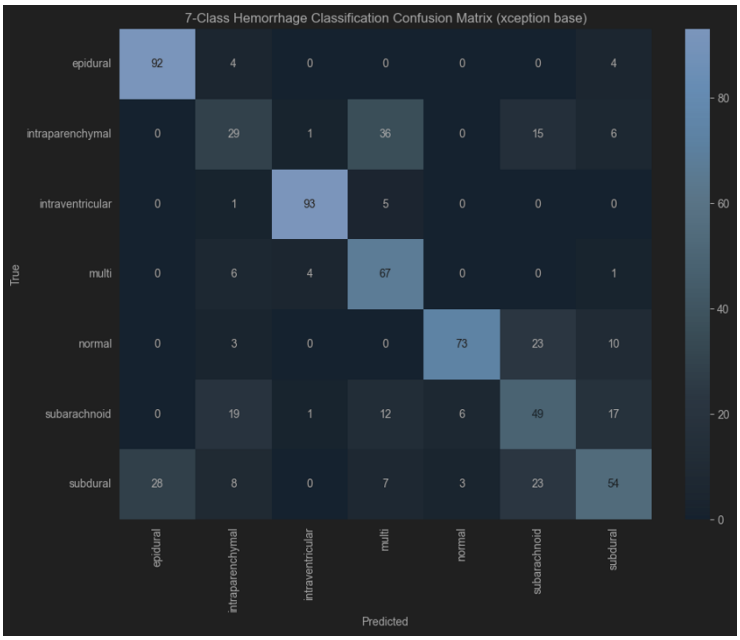

Figure 8: Confusion Matrix of Xception 7-class categorization

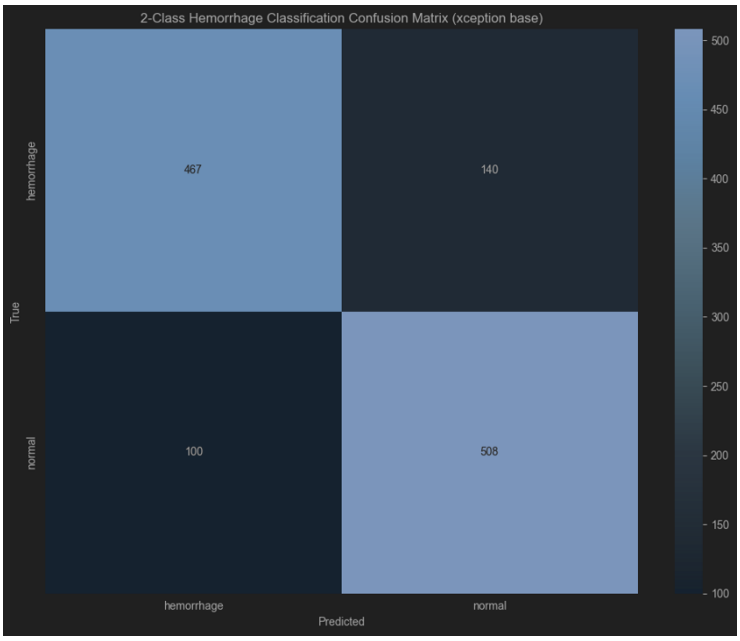

Figure 9: Confusion Matrix of Xception binary categorization

The Xception backbone models show a well-defined diagonal in the confusion matrices, indicating strong overall performance. However, class imbalances persist. Notably, the Xception 7-class model continues to struggle with the intraparenchymal class, frequently misclassifying it as "multi." The ResNet50 model performs even worse in this regard, failing to predict the intraparenchymal class altogether. Since both models exhibit this consistent difficulty, it suggests there may be subtle features of the intraparenchymal class that are inherently challenging for CNNs to learn. Incorporating additional information—such as the maximum contrast window—or applying class-specific augmentation could help improve detection performance for this underrepresented and complex class.

A summary of the accuracies achieved by all models is presented in the table below.

Number of Classes	7	6	2
Basic Model Accuracy	15.02%	-	-
ResNet50 Model Accuracy	30.29%	31.14%	70.54%
Xception Model Accuracy	65.86%	-	80.16%

Application of CNNs for Segmentation

Segmentation masks were available for a subset of the dataset, enabling the training of models to localize hemorrhage regions within CT scans. These masks were manually annotated by trained personnel, who outlined the boundaries of hemorrhages in each image. Custom pre-processing scripts converted the provided boundary coordinates into binary mask images, aligned spatially with the original CT slices.

As in the classification task, the brain, bone, and subdural windows of the CT scans were merged into RGB composite images, while the binary masks served as the ground truth outputs. However, due to the typically small size of hemorrhagic regions relative to the full scan, this task exhibits severe class imbalance—most pixels belong to the non-hemorrhaged background. Consequently, conventional pixel-wise accuracy metrics are misleading, often favoring trivial predictions of entirely non-hemorrhaged regions.

To address this, Dice loss was used as the optimization criterion. This loss function is well-suited for segmentation tasks involving imbalanced data, as it directly measures the overlap between predicted and ground truth masks. The Dice coefficient (also known as the Sørensen–Dice index) is defined as:

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$

Where:

- P is the predicted binary segmentation mask.
- G is the ground truth binary mask.

- $|P \cap G|$ is the number of pixels where both prediction and truth are 1 (the intersection).
- $|P|$ and $|G|$ are the number of positive pixels in the prediction and ground truth, respectively.

In terms of true positives (TP), false positives (FP), and false negatives (FN), the Dice coefficient can also be written as:

$$\text{Dice} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

To use this metric as a loss function, we subtract it from 1:

$$\text{Dice Loss} = 1 - \text{Dice}$$

For differentiability and numerical stability, especially when using soft probability maps from the network, a smoothed version is used:

$$\text{Dice Loss} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}$$

Where $p_i \in [0, 1]$ is the predicted probability for pixel i, $g_i \in \{0, 1\}$ is the true label, and ϵ is a small constant to avoid division by zero.

First, a basic custom U-Net was implemented, but it produced suboptimal results, often predicting either entirely empty masks or overly large hemorrhage regions, despite multiple attempts to refine its structure. As a result, a more robust model was constructed using similar architecture to the CNN used for classification. Specifically, a pre-trained ResNet50 was used as the encoder backbone within a U-Net framework, and the decoder followed the same structural pattern as the classification CNN. An additional upsampling block was added to return the output to the original input resolution of 512×512, followed by a final sigmoid activation layer to produce the binary segmentation mask.

This model was trained for 20 epochs using the Adam optimizer, with ResNet50 weights frozen during training. The resulting test accuracy was 98.94%; however, as previously noted, this metric is misleading due to extreme class imbalance—predicting entirely empty masks could yield a similarly high accuracy. A more meaningful evaluation shows a test precision of 77.32%, indicating that when the model predicts a pixel as hemorrhaged, it is correct 77.32% of the time. The recall was 71.65%, meaning the model correctly identified 71.65% of the hemorrhaged pixels in the ground truth masks. Six randomly selected segmentation results are shown below to illustrate the capabilities of the model.

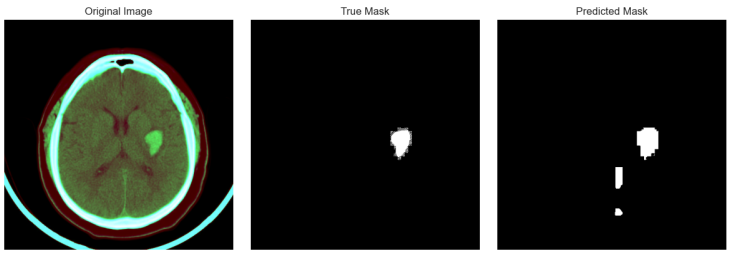

Figure 10: First example image, true mask and generated mask

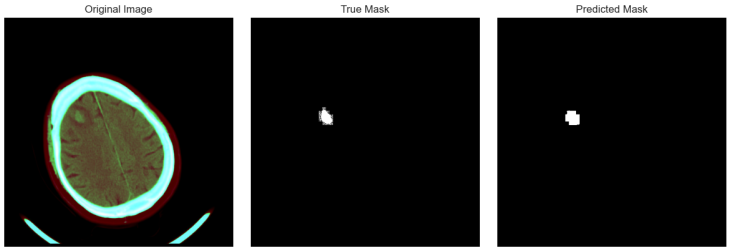

Figure 11: Second example image, true mask and generated mask

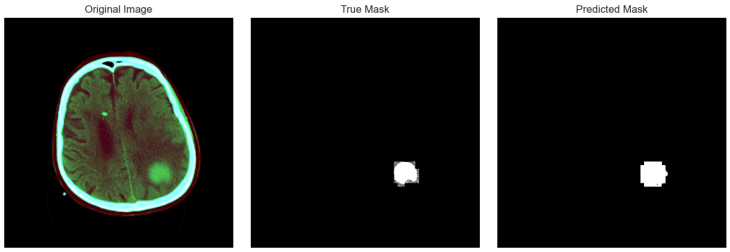

Figure 12: Third example image, true mask and generated mask

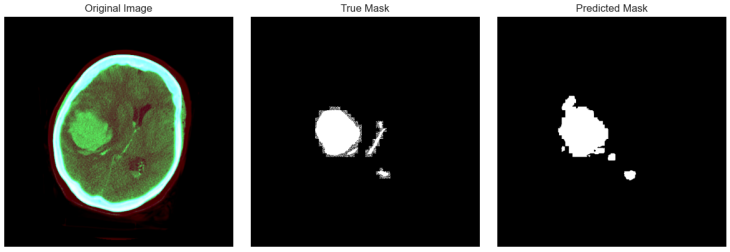

Figure 13: Fourth example image, true mask and generated mask

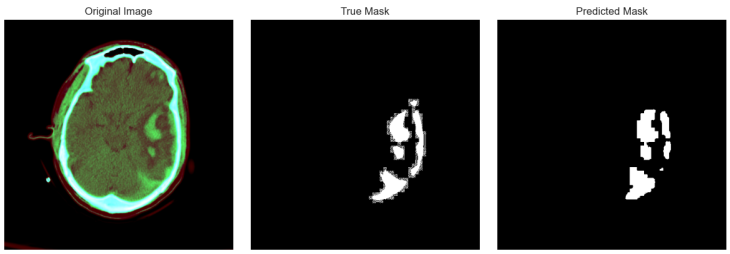

Figure 14: Fifth example image, true mask and generated mask

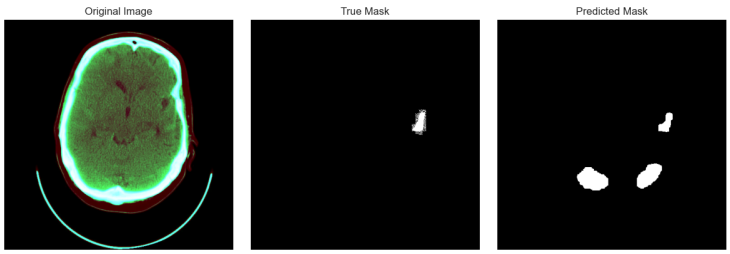

Figure 15: Sixth example image, true mask and generated mask

As shown in the example images, the model performs well in most cases. Of the six examples, only the sixth exhibits a notable false positive, incorrectly highlighting a large non-hemorrhaged region. The first example also includes a small erroneous region. These errors suggest that the model occasionally misidentifies normal regions as significant hemorrhage regions. However, most of the imperfections in the model are in the loss of fine-grain detail. The predicted boundaries can appear coarse and sometimes miss narrow connecting structures, as seen in the fourth and fifth examples. Despite these limitations, the model shows significant potential. It successfully localizes hemorrhage regions, which could be valuable in clinical contexts where approximate location is sufficient—for example, guiding procedures to drain hemorrhages. With further training, it is likely the model's precision and boundary accuracy could improve. However, training time and computational constraints limited the extent of optimization in this iteration. A more granular model could also be built which does not limit the final mask to being binary. Continuous probability masks could be predicted, giving more natural, nuanced boundaries to the regions.

Conclusion

This project evaluated multiple machine learning methodologies for the classification and segmentation of brain hemorrhages on CT scans, comparing classical machine learning techniques with advanced deep learning architectures.

Classical methods—including Logistic Regression, kNN, Decision Trees, Random Forests, QDA, and kernel-based SVMs—consistently underperformed due to their limited ability to model complex, nonlinear relationships within high-dimensional imaging data. Conversely, CNNs, particularly those leveraging transfer learning (ResNet50 and Xception), demonstrated significant improvements in classification tasks. Xception outperformed other models, achieving notable accuracy in both binary (80.16%) and multi-class (65.86%) classifications. Additionally, the ResNet50-based U-Net demonstrated strong capabilities in segmentation tasks, effectively localizing hemorrhage regions despite inherent challenges such as class imbalance and loss of fine-grain detail. Although further optimization is necessary to enhance performance, particularly for challenging hemorrhage subtypes like intraparenchymal hemorrhages, the implemented deep learning approaches substantially advanced automated hemorrhage detection capabilities. Future work should explore more extensive training datasets, advanced augmentation strategies, and refined architectures to further improve the accuracy, robustness, and clinical applicability of these AI tools.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learn-ing for image recognition." 2015.

[2] Chollet, F.: "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2017.

[3] S. Kuo *et al.*, "Deep learning for intracranial hemorrhage detection on CT: A meta-analysis," *Radiology: Artificial Intelligence*, vol. 4, no. 4, pp. e210069, 2022.

[4] G. Yüce *et al.*, "Three-dimensional segmentation of brain hemorrhages using 3D U-Net," *Computers in Biology and Medicine*, vol. 155, 106480, 2023.

[5] Y. Peng *et al.*, "AttFocusNet: Hemorrhage segmentation with attention guidance," *Medical Image Analysis*, vol. 75, 102275, 2022.

[6] J. Lin *et al.*, "Hybrid radiomics and deep learning for hemorrhage classification," *Scientific Reports*, vol. 13, no. 1, 4391, 2023.

[7] L. Chen *et al.*, "Radiomic texture features for ICH subtype classification," *Journal of Medical Imaging*, vol. 10, no. 1, 012004, 2023.

[8] A. Chilamkurthy *et al.*, "Development and validation of deep learning algorithms for detection of critical findings in head CT scans," *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.