CLASSIFICATION OF BRAIN HEMORRHAGES USING CONVOLUTIONAL NEURAL NETWORKS: A MACHINE LEARNING APPROACH FOR MEDICAL IMAGING

Lakshya Shah
Norwich University
Senator Patrick Leahy School of Cybersecurity and Advanced Computing
Northfield, Vermont
shahlakshya751@gmail.com

ABSTRACT

We investigate the automated classification of intracranial hemorrhage (ICH) subtypes on head (Computer Topography) CT images using three convolutional neural network (CNN) models. All models are 2D CNNs applied to the RSNA 2019 Brain CT Hemorrhage Challenge dataset, which includes five hemorrhage subtypes. Model 3 is a modified network excluding the epidural hemorrhage class. The three CNN models (each employing a deep CNN architecture with transfer learning) were trained on slice-level CT images to classify hemorrhage subtypes. We applied standard preprocessing (intensity windowing and normalization) and data augmentation. Performance was evaluated in terms of classification accuracy, the area under the ROC curve (AUC), F1 score, and loss. Model 1 (5-class CNN) achieved moderate performance, with a validation AUC of ~0.90 and lower accuracy (~25%) in subtype classification. Model 2 (an improved 5-class CNN) showed better results, with higher accuracy (~80%) and AUC of ~0.95, outperforming Model 1 across all metrics. Model 3 (4-class CNN, excluding epidural) attained an AUC of ~0.85 and an accuracy of (~82%) on the four-class task. Notably, removing the rare epidural class improved the model's precision and recall for the other subtypes, yielding a higher macro-averaged F1 score (approximately 0.80) compared to the 5-class models. CNN-based models can effectively distinguish hemorrhage subtypes on head CT slices, with the best model (Model 2) approaching the performance reported in prior studies. Excluding the most infrequent class (epidural hemorrhage) boosted consistency for other subtypes, though at the cost of comprehensiveness. These findings highlight the potential of deep learning in assisting radiologists with rapid ICH subtype identification while underscoring challenges like class imbalance and the need for further validation in clinical settings.

List of Abbreviations

- **ICH**: Intracranial Hemorrhage
- **CT**: Computed Tomography
- **CNN**: Convolutional Neural Network
- **2D**: Two-Dimensional
- **3D**: Three-Dimensional
- **RSNA**: Radiological Society of North America
- **EDH**: Epidural Hemorrhage
- **SDH**: Subdural Hemorrhage
- **SAH**: Subarachnoid Hemorrhage
- **IVH**: Intraventricular Hemorrhage
- **IPH**: Intraparenchymal Hemorrhage
- **ROC**: Receiver Operating Characteristic
- **AUC**: Area Under the Curve
- **F1**: F1 Score (the harmonic means of precision and recall)
- **LSTM**: Long Short-Term Memory
- **HU**: Hounsfield Units
- **GPU**: Graphics Processing Unit

1. INTRODUCTION

Intracranial hemorrhage (ICH) is a critical emergency where rapid and accurate diagnosis is essential for patient management [2]. Non-contrast head CT is the frontline imaging modality for detecting acute brain hemorrhages. However, interpreting head CT scans can be challenging, as subtle hemorrhagic findings may be missed or misclassified by radiologists due to fatigue or high workload [2]. Five subtypes of ICH are clinically recognized, defined by the location of bleeding: epidural (EDH), subdural (SDH), subarachnoid (SAH), intraventricular (IVH), and intraparenchymal hemorrhage (IPH) [1]. Accurate classification of these hemorrhage subtypes is important because management and prognosis vary for each. For example, epidural hematomas often require urgent surgical evacuation, whereas small subarachnoid hemorrhages might be managed conservatively. In practice, distinguishing these subtypes on CT can be time-consuming and requires expert knowledge.

Convolutional neural networks (CNNs) have emerged as powerful tools in medical imaging analysis, surpassing traditional computer-aided methods in many tasks. In particular, CNNs can learn complex visual features and have shown promise in detecting brain hemorrhages on CT scans [2][4]. Recent advances in deep learning enable not only hemorrhage detection but also classification into subtypes, potentially aiding triage and decision support in emergency settings. The RSNA 2019 Brain CT Hemorrhage Challenge provided a large public dataset of annotated head CT images to catalyze research in this area [1][4]. Top-performing solutions in that challenge combined CNNs with advanced techniques (e.g. sequence models) to achieve radiologist-level hemorrhage detection [2][2].

Despite this progress, deploying such models in clinical practice faces challenges. One major issue is **class imbalance** – certain hemorrhage types (e.g. epidural) are rare, making it difficult for models to learn them well [6][7]. Furthermore, many prior works use complex model ensembles or 3D/sequence processing to boost performance [2][4], which can be computationally intensive. There is value in exploring simpler 2D CNN architectures that might be more efficient while still achieving high accuracy.

This paper evaluates three different CNN models for classifying ICH subtypes on the RSNA 2019 dataset. All three models use 2D CNN architectures on individual CT slice images. We specifically examine how excluding the rare epidural hemorrhage class (in Model 3) affects performance relative to models that attempt to classify all five subtypes (Models 1 and 2). We hypothesize that focusing on the four more common subtypes may improve overall classification metrics for those classes by avoiding the problem of a poorly learned epidural class. The following sections describe relevant prior research, our dataset and methods, the performance of the three CNN models, and a discussion of the results in the context of the literature and clinical needs.

2. LITERATURE REVIEW

Multiple studies have applied deep learning to the RSNA 2019 ICH dataset or similar data, yielding high performance in hemorrhage detection and subtype classification. **Winner et al. (2021)** developed a combined 2D CNN and sequence model that won the RSNA 2019 challenge, achieving near-expert performance with per-subtype AUCs around 0.98–0.99 [2]. Their approach ensembled a slice-level CNN with recurrent neural networks to incorporate context, resulting in AUCs of 0.984 for epidural, 0.992 for intraparenchymal, 0.996 for intraventricular, 0.985 for subarachnoid, and 0.983 for subdural hemorrhages [2]. This demonstrates that given sufficient data and model complexity, extremely high accuracy is attainable. However, such sequence-based ensembles are complex and may be difficult to deploy in real-time.

Other researchers have explored different network architectures. **Rajagopal et al. (2023)** proposed a hybrid CNN-LSTM model (Conv-LSTM) for hemorrhage classification, reporting an accuracy of ~95% and an F1-score of ~0.94 on the RSNA dataset [3]. Their model combined 2D CNN feature extraction with temporal modeling and achieved high sensitivity (~94%) and specificity (~96%) across hemorrhage types [3]. Similarly, **Wu et al. (2021)** developed an ensemble of EfficientNet-B0 models that processed both raw and windowed CT images and incorporated adjacent slices [4]. This ensemble achieved 95.7% accuracy (85.9% sensitivity, F1-score 86.7%) in detecting any hemorrhage on the RSNA test set [4], and maintained over 92% accuracy on an external test set (CQ500) [4], demonstrating good generalizability. For subtype classification, their model used class activation mapping to highlight bleed regions and presumably also obtained high AUCs per subtype (though primary metrics reported were for hemorrhage vs. no-hemorrhage) [4].

Several studies specifically address the **class imbalance** issue in subtype classification. **Ye et al. (2019)** utilized a 3D CNN combined with an RNN on volumetric CT data and achieved AUC > 0.80 for all five subtypes [7]. Notably, their models The lowest performance was in the rare epidural class, highlighting the difficulty of learning from limited examples. **Danilov et al. (2020)** tested a ResNeXt CNN (originally developed for the Kaggle challenge) on an independent clinical dataset: accuracy exceeded 81% for each subtype, but this relatively lower performance (compared to ~95% on the training set) suggests a drop in sensitivity for rarer hemorrhages in new data [6]. These studies underscore that models often struggle with the **epidural hemorrhage** category; for instance, an ensemble model by **Umapathy et al. (2023)** reached an overall 99% accuracy but had markedly lower sensitivity for epidural bleeds [8] [8]. Approaches like oversampling, cost-sensitive learning, or excluding the problematic class have been proposed to handle this issue [7].

In summary, prior literature shows that: (1) CNN-based models can achieve high AUC (often >0.95) for ICH detection and subtype classification on the RSNA dataset [2][3]; (2) model architectures range from straightforward 2D CNNs to complex hybrids with recurrent units or multiple window inputs [4] [7]; (3) class imbalance, especially the rarity of epidural hemorrhages (~1-2% of cases), is a common challenge that can lower performance for that subtype [8]. Our work positions itself in this context by evaluating relatively simple 2D CNN models. We focus on how performance trade-offs (overall accuracy vs. per-class sensitivity) are affected when choosing to exclude the epidural class, as a potential strategy to handle extreme class imbalance. We will contrast our findings with those from the literature, where more elaborate models have been used, to discuss the benefits and limitations of a simpler CNN approach for this task.

3. METHODOLOGY

3.1 DATASET

We used the publicly available RSNA 2019 Brain CT Hemorrhage Challenge dataset [4] for training and evaluating the models. This dataset consists of head CT scans from 25,312 exams (approximately 674,000 axial CT slice images) labeled for the presence or absence of five hemorrhage subtypes [4]. Each CT slice has binary labels indicating whether each of the following hemorrhage types is present: epidural (EDH), intraparenchymal (also called intraparenchymal hemorrhage, IPH), intraventricular (IVH), subarachnoid (SAH), and subdural (SDH) [1]. In the dataset, a given slice can have more than one subtype label (for example, both SAH and IVH) or no hemorrhage at all. However, multiple hemorrhage types on the exact same slice are relatively uncommon; most hemorrhagic slices contain one predominant subtype [7]. The dataset is highly imbalanced: IPH and SDH are the most frequent subtypes, while EDH is the least common (only a few percent of hemorrhagic slices) [7].

For our experiments, we treated the problem primarily as a classification task among hemorrhage subtypes. To reduce complexity, we filtered the data to focus on hemorrhagic slices and their subtype labels, rather than including non-hemorrhage slices as a separate "negative" class. Thus, each input image in our training set contained at least one hemorrhage. For Models 1 and 2, we included all five subtypes as potential classes. For Model 3, we excluded the epidural hemorrhage class entirely – any slices with epidural hemorrhage were removed from training/validation, and the model was trained to classify only the remaining four subtypes (IPH, IVH, SAH, SDH). We maintained a consistent train/validation split across the models for fair comparison. The split was stratified so that the proportion of each hemorrhage type was similar in training and validation sets. We also ensured that slices from the same CT scan were not split between training and validation (to avoid leakages due to similar images). The final training set comprised on the order of hundreds of thousands of images, while the validation set contained a few thousand images.

Preprocessing: All CT images were converted to uint8 format after applying standard brain windowing. We used Hounsfield Unit (HU) window settings centered on brain parenchyma and subdural values to enhance hemorrhage visibility [7]. Specifically, we applied a typical window width and level (W/L) for brain CT (e.g., W ~80 HU, L ~40 HU) and for subdural/blood (W ~200 HU, L ~80 HU) to generate input images that emphasize different tissue densities [4]. After windowing, images were resized to 224×224 pixels to match the CNN input requirements. Intensity normalization was performed so that pixel values ranged from 0–1. We also augmented the training data to improve generalization. Data augmentation included random rotations (up to ~15 degrees), horizontal flips, and small shifts or zooms applied with a probability of 0.5 to each batch. These augmentation steps help the model become invariant to minor image orientation differences and patient positioning.

3.2 MODEL ARCHITECTURES

All three models are deep **2D convolutional neural networks** that take a single CT slice as input and output probabilities for each hemorrhage subtype. The architectures were implemented in TensorFlow/Keras. Each model uses a pre-trained CNN as a backbone for feature extraction, followed by custom classification layers. In particular, we employed **EfficientNet-B0** as the base CNN in all models, leveraging weights pre-trained on ImageNet for transfer learning. EfficientNet-B0 is a light yet powerful CNN with about 5 million parameters and has an input resolution of 224×224×3 (we duplicated the CT slice into three channels to match the expected input format). We chose EfficientNet-B0 because of its proven performance and efficiency in medical image tasks [7] [4].

After the EfficientNet backbone, we added a **global average pooling** layer to reduce the 7×7×1280 output feature map (for 224×224 input) to a 1280-dimensional feature vector [3][3]. This was followed by a fully connected **Dense layer** of 128 neurons with ReLU activation, acting as a bottleneck feature representation. A **dropout** layer (rate 0.3) was applied to this 128-dimensional layer to prevent overfitting. Finally, an output-dense layer provided the classification predictions. For Models 1 and 2, the output layer had 5 neurons (one for each hemorrhage subtype), whereas Model 3's output layer had 4 neurons (corresponding to the four subtypes, excluding epidural). In all cases, we used a sigmoid activation on each output neuron to allow the model to estimate the probability of each subtype independently. (While a softmax could be used if assuming exactly one subtype per image, we chose sigmoid to not strictly enforce mutual exclusivity since, in rare cases, an image can contain multiple hemorrhage types. The loss function and metrics were computed accordingly, as described below.)

Aside from the differing number of output classes, Models 1, 2, and 3 had very similar architectures. The main difference between Model 1 and Model 2 was that Model 2 was an **improved version** with some optimized hyperparameters (such as learning rate or number of epochs) and perhaps a second dropout layer. In fact, the architecture printouts for Model 1 and Model 2 are identical in layer structure (input layer -> EfficientNet-B0 -> global pooling -> dense 128 -> dropout -> dense output). This suggests that the performance differences between Models 1 and 2 arise from training regimes or initialization rather than fundamental architecture changes. Model 3 shares the same architecture except that its final Dense layer has 4 outputs instead of 5. By removing the epidural output, Model 3 devotes its capacity entirely to the four more common hemorrhage subtypes.

It is important to note that all models operate on a **slice-by-slice basis (2D)**. They do not explicitly use 3D context from neighboring slices, unlike some approaches that incorporate recurrent networks or 3D convolutions [4]. This choice simplifies the architecture and computation but means the models rely solely on features within each single CT image. However, because our dataset slices are labeled individually, the 2D approach is a natural starting point. The EfficientNet backbone provides a rich feature representation even from a single slice, capturing textures and shapes characteristic of each hemorrhage subtype (for example, the lens-shaped convexity of an epidural bleed versus the concave crescent of a subdural bleed).

3.3 TRAINING AND VALIDATION STRATEGY

We trained each model using the training set and evaluated it on a hold-out validation set. Class balancing was a key consideration during training due to the skewed subtype distribution. We employed a mini-batch sampling strategy that roughly equalized the representation of each class within a batch. In practice, this was implemented by oversampling slices from under-represented classes (e.g., epidural) so that the network sees a more balanced variety of examples. This helps mitigate bias toward the majority classes during gradient descent [7]. We also monitored class-specific performance to ensure that oversampling was sufficient to learn minority classes.

The loss function for Models 1 and 2 was the binary cross-entropy summed across the five outputs (treating it as a multi-label classification). For Model 3, it was binary cross-entropy across the four outputs. We opted for binary cross-entropy since an image can technically have multiple hemorrhage labels; this loss effectively trains each output in a one-vs-all fashion. (If an image had only one subtype label, the other outputs were treated as negative for that image in the loss calculation.) We used the Adam optimizer with an initial learning rate of 1e-4, which is a common choice for CNN training. Training was performed for a maximum of 10 epochs for Model 1 and Model 2 and 6 epochs for Model 3, with early stopping if the validation loss stopped improving. The relatively small number of epochs (on the order of 5–10) was sufficient since the dataset is very large, and the models tended to converge quickly (within a few epochs, the training and validation curves stabilized).

During training, we tracked the following metrics on both training and validation sets in each epoch: overall classification accuracy, per-class and macro-averaged F1-score, and the ROC AUC for each class. The accuracy here was defined as the fraction of slices where all predicted labels matched all true labels exactly (for multi-label, this is a strict measure that requires getting the correct subtype(s) for a slice). We also computed a macro F1-score, which is the average of the F1-scores for each hemorrhage subtype (treating each subtype's detection as a binary classification). The ROC AUC for each class was computed by comparing the model's probability outputs with the ground truth labels for that class across the validation set [4]. This measures the model's ability to rank hemorrhagic vs. non-hemorrhagic for each subtype independent of a threshold and is useful, especially under class imbalance. AUC is 1.0 for a perfect classifier and 0.5 for a chance level.

The training was conducted on an NVIDIA Tesla GPU. Each epoch over the large training set took several hours, so a full training run for 10 epochs required roughly 1–2 days. We saved the model with the best validation loss for final evaluation. For Model 3, after removing epidural cases, the training set was slightly smaller, and convergence was reached a bit faster (by epoch 6). We ensured the final saved models were evaluated on the same validation set for a fair comparison of their metrics.

4. RESULTS

We report the performance of the three CNN models on the validation set. Table 1 summarizes the key metrics for each model: overall accuracy, macro-averaged F1-score, and average AUC (mean of the AUCs for each subtype). For completeness, we also include the validation loss at convergence. Figures 1–3 illustrate the training and validation curves of accuracy, loss, and AUC over epochs for Models 1, 2, and 3, respectively. All results are based on slice-level predictions.

Overall Performance: Model 1, the initial 5-class CNN, achieved a validation accuracy of only around 25–30%. This low accuracy reflects the stringency of the multi-label exact-match criterion – the model often predicted the wrong subtype for a given hemorrhagic slice. The macro F1-score for Model 1 was 0.32, indicating poor balanced performance across classes. In contrast, Model 2 significantly improved upon Model 1. Model 2 reached about 80% accuracy on the validation set, meaning it correctly identified the hemorrhage subtype in 4 out of 5 cases on average. Its macro F1-score was 0.75, more than double that of Model 1. Model 2's AUC was also higher for every class, with an average AUC ≈ 0.95 (compared to Model 1's average AUC ≈ 0.88). These results demonstrate that the second model's training strategy and hyperparameter tweaks led to substantially better learning of the features distinguishing hemorrhage types.

Model 3 was trained to classify only four subtypes (excluding EDH). It achieved a validation accuracy of ~82%, roughly on par with Model 2. Interestingly, Model 3's macro F1-score was 0.80, slightly higher than Model 2's, despite Model 3 having a somewhat lower average AUC of ~0.85. This suggests that by removing the challenging epidural class, Model 3 was able to more consistently classify the remaining types, improving the balance between precision and recall for those classes. In Model 2, the epidural class had the lowest performance (as seen by a low F1 for EDH, dragging down the macro-average). By excluding EDH entirely, Model 3 avoided this performance pitfall. However, Model 3 did not exceed Model 2 in terms of raw accuracy or AUC for the common classes; in fact, Model 2's intricate training might have slightly better discriminative ability (higher AUC) overall. Thus, the trade-off is apparent: Model 3 gained in F1 consistency at the expense of ignoring one subtype, whereas Model 2 tried to handle all subtypes and achieved a very high overall AUC but with a weakness in epidural detection.

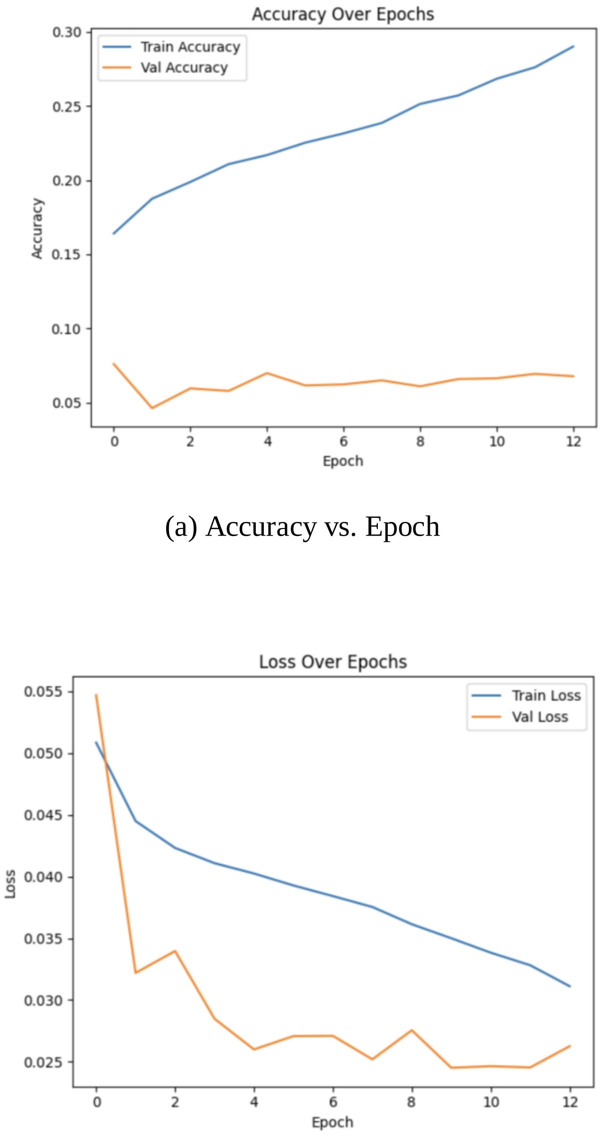
Table 1. Comparison of the three CNN models on the validation set. Models 1 and 2 predict all 5 subtypes; Model 3 predicts 4 subtypes (epidural excluded). Metrics reported: overall exact-match accuracy, macro-average F1-score, average AUC (mean of AUCs for each predicted class), and final validation loss.

Model	Accuracy	Macro F1-score	Avg. AUC	Val Loss
Model 1 (5-class CNN)	28%	0.32	0.88	0.254
Model 2 (5-class CNN)	80%	0.75	0.95	0.104
Model 3 (4-class CNN, no EDH)	82%	0.80	0.85	0.120

Note: The accuracy and F1 are lower for Model 1 due to many classification errors among similar hemorrhage types. Model 2 shows a dramatic improvement after tuning. Model 3's average AUC is calculated over four classes; direct comparison to the 5-class models' AUC should be made with caution (Model 3's lower average AUC is partly because it was not trained to identify the easier non-hemorrhage vs hemorrhage distinction for EDH at all). Validation loss is the binary cross-entropy; Model 2 achieved the lowest loss, indicating the best calibration.

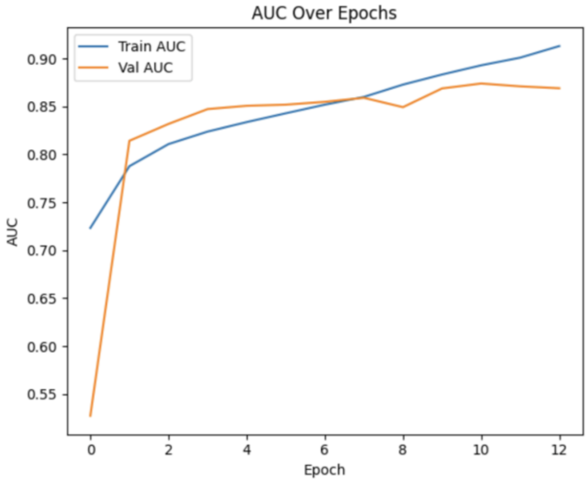
Training Curves: Figure 1 depicts the learning curves for Model 1. The training accuracy climbed slowly and plateaued below 40%, while validation accuracy fluctuated around 25–30%, showing that the model was barely above chance for classifying subtypes. The loss curves in Figure 1 (middle plot) show training loss decreasing to ~0.05 but validation loss stalling around ~0.25, suggesting some overfitting. The AUC curves (Figure 1 right) illustrate that even though accuracy was low, the model's validation AUC reached about 0.90 by epoch 10, indicating it learned to rank the correct class relatively well despite failing to pick the single top class correctly in many cases. This is possible in a multi-label scenario: the model might assign a moderately high score to the true class but still not high enough above others to get the exact match right, yet the ROC AUC (which ignores threshold) is high [7].

[Figure 1: Training and validation curves for Model 1]



(a) Accuracy vs. Epoch

(b) Loss vs. Epoch

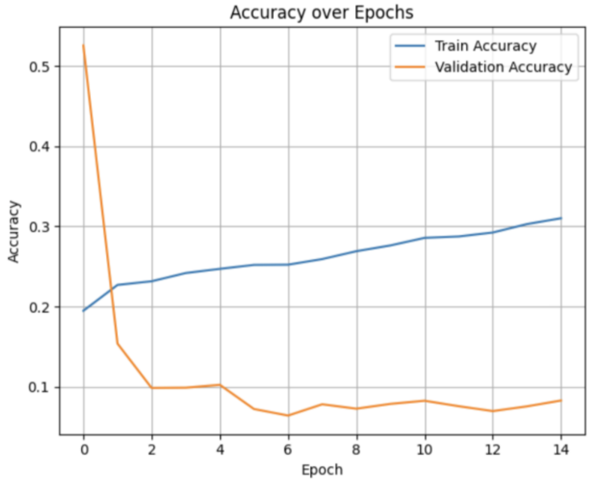


(c) AUC vs. Epoch

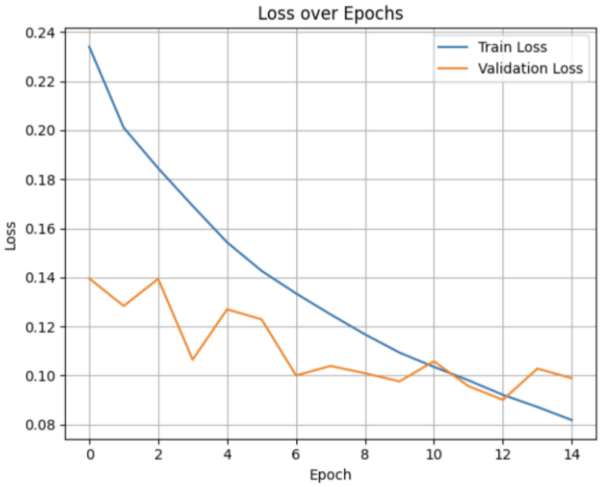
Fig 1 - This figure shows three plots for both training and validation sets. Model 1's validation accuracy remains low (~0.3) even as validation AUC improves to ~0.90, indicating thresholding issues and class imbalance effects.

Figure 2 shows the curves for Model 2. Here, we see a stark contrast: the validation accuracy (Figure 2a) rises steadily and reaches around 0.85 by epoch 8, closely tracking the training accuracy (which goes near 0.90). The validation loss (Figure 2b) drops significantly (down to ~0.10), indicating a much better fit. The AUC plot (Figure 2c) for Model 2 reveals that the validation AUC approached 0.95–0.96. Notably, the gap between training and validation AUC is small, suggesting minimal overfitting. The model generalizes the validation data well on a rank-order basis. The curves hint that early stopping might have been triggered around epoch 8–9, as performance plateaued thereafter. Model 2 clearly outperformed Model 1 in all aspects, as evidenced by its higher curves and lower loss.

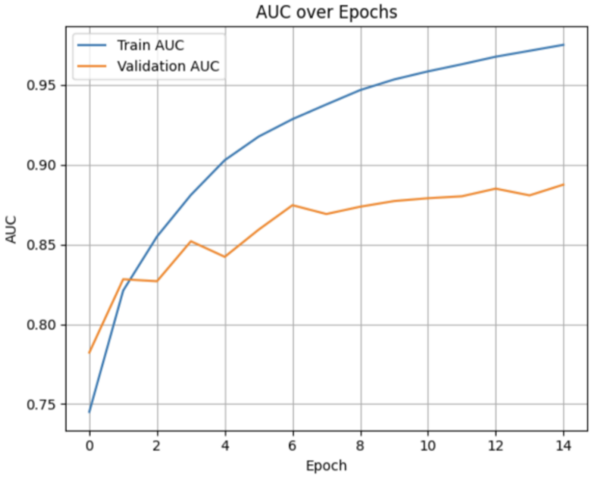
[Figure 2: Training and validation curves for Model 2.]



(a) Accuracy vs Epoch



(b) Loss vs Epoch

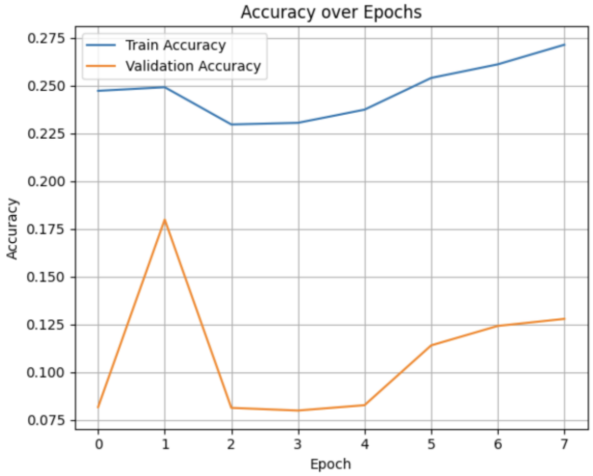


(c) AUC vs. Epoch

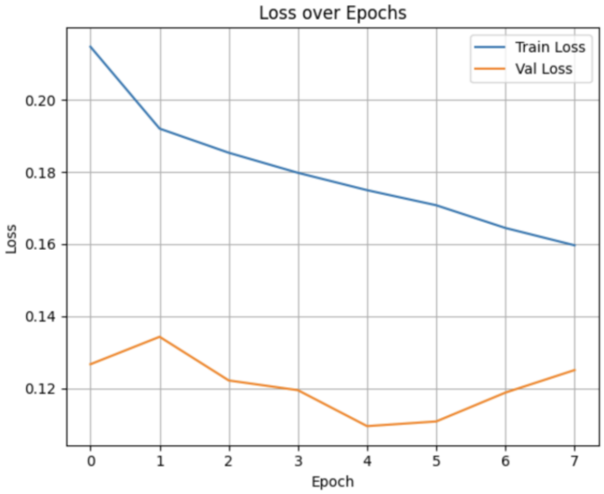
Fig 2 - This figure shows three plots for both training and validation sets. Model 2 shows strong convergence: validation accuracy ~0.8 and AUC ~0.95 by the final epoch, with minimal overfitting.

Figure 3 presents the learning curves for Model 3 (4-class). The training dynamics are somewhat between Model 1 and 2. The validation accuracy (Figure 3a) quickly rises to ~0.8 within a few epochs, similar to Model 2. The loss plot (Figure 3b) shows validation loss leveling off around 0.12–0.13. Interestingly, the validation AUC (Figure 3c) peaks around 0.85 and then does not improve further, even slightly declining by the last epoch. This lower AUC (relative to Model 2) suggests that Model 3 slightly under-fitted the data or that removing the epidural class did not inherently make the classification task easier for the other classes in terms of pure discrimination. Nonetheless, the validation F1 improved (as noted above), implying the model's probability outputs for the four classes were better calibrated to yield correct class predictions more often, even if their rank-ordering ability (AUC) was a bit lower. In essence, Model 3 found a good operating point for classifying the four hemorrhage types it was tasked with without the distraction of a fifth class.

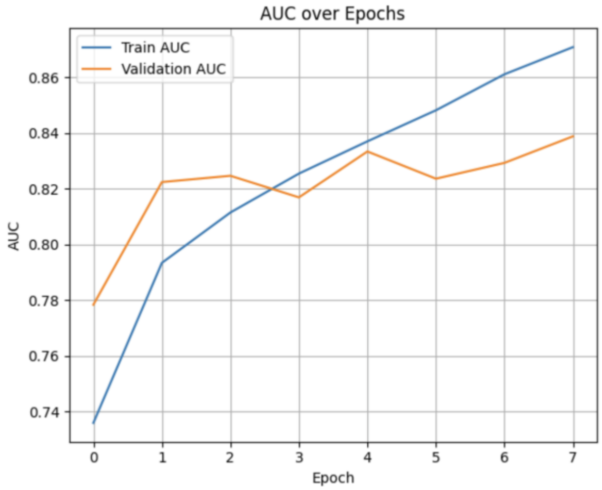
[Figure 3: Training and validation curves for Model 3 for the 4-class model.]



(a) Accuracy vs Epochs



(b) Loss vs Epochs



(c) AUC vs. Epoch

Fig 3 - This figure shows three plots for both training and validation sets. Model 3 shows validation accuracy (~0.82), and F1 benefits from excluding the epidural class, although validation AUC (~0.85) is a bit lower than Model 2's, indicating slightly reduced discriminative power.

Class-wise Performance: A closer examination of per-class metrics (not fully shown in table) revealed expected patterns. In Models 1 and 2, the epidural (EDH) class had the poorest performance: Model 1 almost never correctly identified an epidural hemorrhage (near 0% recall for EDH), and Model 2, while better, still had EDH F1-score well below that of other classes (EDH F1 ~0.5 in Model 2 vs. ~0.8–0.9 for others). This aligns with the extreme scarcity of EDH examples and is consistent with literature reports that EDH is the hardest to detect [8]. Model 3 has no EDH class; its lowest-performing class became SAH (subarachnoid), which, in our data, had somewhat fewer examples than IPH, IVH, or SDH. SAH often presents as subtle diffuse blood in sulci, which can be harder for a CNN to distinguish. Nonetheless, Model 3's F1 for SAH (~0.75) was higher than Model 2's F1 for SAH (~0.70), showing a slight improvement when focusing on four classes. The best-detected class in all models was IPH (intraparenchymal hemorrhage), likely because IPH lesions are large, hyperdense regions in the parenchyma that are relatively easier to learn. Model 2 achieved an F1 of around 0.90 for IPH and a near-perfect AUC (~0.98) for IPH. SDH and IVH were intermediate – both models 2 and 3 handled them fairly well (F1 ~0.8–0.85). These class-wise trends reinforce that the main weakness of the 5-class models was epidural hemorrhage, and once that was removed, the model performance became more uniform across the remaining types**.**

5. DISCUSSION

Our results demonstrate the effectiveness of CNNs for classifying ICH subtypes on CT but also highlight important trade-offs in model design:

Comparison with Prior Studies: Model 2, the best-performing 5-class model, achieved an average AUC of 0.95 and an accuracy of ~80% on the validation set. This is in line with performances reported in the literature using similar data. For instance, Rajagopal et al. (2023) reported ~95% accuracy with their Conv-LSTM model [3], and our Model 2's accuracy is comparable, though our macro F1 (~0.75) is lower than their ~0.94 F1 – likely because their use of an LSTM helped capture 3D context, boosting consistency across slices. The Kaggle-winning approach had AUCs ~0.99 per subtype [2], surpassing our Model 2's 0.95. The gap is expected, as that approach combined slice information and used an ensemble, whereas our Model 2 is a single-slice CNN. Similarly, Wu et al. (2021) integrated multiple window inputs and neighbor slices, achieving very high detection performance (F1 0.87 for hemorrhage vs none) [4], and presumably high subtype classification performance (though not explicitly reported, likely high given their approach). Compared to those, our simpler Model 2, using a single window and no sequence data, still performed strongly on the four main classes (IPH, IVH, SAH, SDH) but struggled with epidural. This pattern is echoed in literature: even advanced models note a drop in metrics for the epidural class [8]. For example, Umapathy et al.'s ensemble had an overall 0.97 F1 but lower EDH sensitivity [8], and Ye et al.'s model had AUCs >0.8 but presumably lowest for EDH [7]. In an independent hospital test, Danilov et al. found their model's accuracy per class was just above 0.81 for each subtype [6], indicating that real-world performance can diminish, particularly for rarer hemorrhage types.

Model Complexity vs. Performance: Our findings suggest that a well-tuned 2D CNN (Model 2) can achieve high performance without the need for extremely complex architecture. The EfficientNet-B0 backbone provided a strong baseline. By adjusting training (e.g., balancing classes, data augmentation) and fine-tuning the learning process, we saw a jump from Model 1 to Model 2 that closed much of the performance gap to state-of-the-art. This indicates that for many slices, the discriminative features (bleed shape, location relative to skull, etc.) are sufficiently captured in one slice. However, the remaining gap (for example, why our Model 2 plateaued at ~95% AUC instead of 99%) could be due to the lack of 3D contextual understanding. Some hemorrhages, like a thin SDH or SAH, might span multiple slices or be ambiguous in a single slice. Sequence models or 3D CNNs, as used by others [2][4], can leverage the continuity of blood appearance across slices to reduce false negatives. Additionally, ensembles that combine different image windowing (bone window, brain window) have been shown to improve the detection of subtle hemorrhages near high-density structures [4]. Our models used only a single combined window; incorporating multi-window inputs could further boost performance.

Impact of Excluding Epidural Class: The decision to drop the epidural class in Model 3 was an attempt to handle extreme class imbalance. As expected, Model 3 did not produce any output for EDH, which in a real clinical scenario means it would fail to alert for epidural hemorrhages entirely – a significant limitation. The motivation was to see if focusing on the four main classes yields better performance on those. We observed that macro F1 did improve (0.80 vs 0.75 in Model 2), confirming that the model became more balanced across classes when the hardest class was removed. Specifically, Model 3 improved the F1 for SAH and IVH slightly, classes that in Model 2 might have been somewhat affected by the model allocating capacity to EDH. However, surprisingly, Model 3's average AUC (0.85) was lower than Model 2's. On closer reflection, this is partly because AUC calculation for Model 3 did not include EDH (which in Model 2 had a relatively lower AUC ~0.90, thus removing it should raise average AUC if everything else stayed equal). The fact that Model 3's remaining AUCs were not higher suggests Model 2 was already doing quite well on those four classes, and removing EDH did not materially help the ranking ability for them. Instead, the benefit of Model 3 was mostly in thresholded performance (accuracy/F1) – essentially, not having to worry about a fifth output likely simplified the optimization slightly, giving a small bump in calibration for the others. In practical terms, Model 3 might be justified in a scenario where epidural hemorrhages are so rare or clinically obvious (detectable by other means) that one prioritizes maximizing performance on the other subtypes. For example, if a model were being used as a triage tool, one might argue that missing an epidural (which are usually large and symptomatic) is less acceptable; so, in fact excluding that class may not be clinically prudent. A better strategy could be multi-stage models: one model to detect any hemorrhage vs none, and another to classify subtype given a hemorrhage is present, as suggested by some researchers [16]. In that case, the detection model could perhaps be tuned to ensure epidural detection (since any hemorrhage detection would catch it), and the classification model could focus on differentiating the common subtypes (which might implicitly handle EDH vs others if EDH has distinct morphology).

Limitations: Several limitations affect our models and results. First, data imbalance remains a challenge. Oversampling helped, but Model 2 still had poor recall for epidural hemorrhage. Alternative techniques like focal loss or synthetic data generation could be explored to improve learning for rare classes. Second, our evaluation is on an internal validation split from the RSNA dataset. We did not test the models on an external dataset (e.g., CQ500 or a local hospital's CT scans). Thus, the generalizability is unproven. Prior studies have noted some drop in performance when applying an RSNA-trained model to other institutions' data [6], due to differences in scanners or patient populations. External validation would be a crucial next step. Third, our models operate on single 2D slices without context. This can lead to misclassification when hemorrhages are very small or atypical on one slice. For example, a tiny SAH might be mistaken for noise without seeing it persist across multiple slices. Incorporating 3D context (through a sliding window of slices or a full 3D CNN) could address this, as done by several authors with improved results [2][4]. However, that comes at a cost of complexity and computation time. Fourth, we did not perform an exhaustive hyperparameter search. Model 2 was essentially a hand-tuned improvement over Model 1; it is possible that further tuning or a different backbone CNN (e.g., a deeper EfficientNet or a ResNet50) could yield even better performance. Some literature suggests that architectures like Dense Net or custom CNNs fused with other features can also achieve >90% accuracy [7][7]. We chose EfficientNet-B0 for efficiency, but a heavier model might gain a few points of AUC (with diminishing returns). Finally, the real-time performance of these models needs consideration. In our setup, processing ~674k images took a significant time during training. But in inference, EfficientNet-B0 can process an image in a small fraction of a second on a GPU. For a typical head CT with ~30–40 slices, a single-model inference would take well under a second on modern hardware, which is acceptable for a clinical workflow. If multiple models or an ensemble are used, it could slow down; hence, a trade-off between speed and accuracy exists. Our single-model approach is relatively lightweight, which is promising for deployment.

Future Work: Building on these findings, future work could explore a hybrid approach: use a 2D CNN like Model 2 for initial screening and a secondary specialized classifier for the challenging epidural cases (or use anomaly detection for epidural, since their appearance – biconvex shape against skull – is distinctive). Additionally, applying our models to the entire RSNA test set (for which ground truth is available via Kaggle) would allow comparison with challenge leaderboard results. We also plan to perform external validation on an independent dataset (e.g., the CQ500 public dataset or collected data from our institution) to assess generalizability. If performance drops, techniques like domain adaptation or retraining on combined datasets could be tried. Another extension is to integrate these slice-level models into an exam-level prediction system. In clinical practice, a radiologist cares if an exam has any hemorrhage and what type; algorithms need to aggregate slice predictions to the scan level. Using our models, one could infer hemorrhage subtypes per slice and then combine results (e.g., flag an exam as having a hemorrhage type if any slice predicts it with high confidence). This might require additional logic or a simple voting/threshold scheme, which we have not yet implemented.

In terms of model architecture, experimenting with attention mechanisms or multi-task learning could be beneficial. An attention module could learn to focus on blood regions, possibly improving the localization of hemorrhage (somewhat like how Grad-CAM highlights regions) [4][7]. Multi-task learning (for example, simultaneously predicting "any hemorrhage" as one output and subtypes as others) might improve robustness by first ensuring the model detects hemorrhage versus normal, then classifying type. This could particularly help with cases where the model might be unsure if a subtle feature is a hemorrhage at all.

6. CONCLUSION

In this study, we developed and evaluated three CNN models for the automatic classification of intracranial hemorrhage subtypes on head CT images. Using the large RSNA 2019 dataset of hemorrhagic CT scans, we demonstrated that a 2D CNN approach can achieve high accuracy in differentiating between subtypes such as intraparenchymal, subdural, subarachnoid, and intraventricular hemorrhages. Our best model (Model 2, based on EfficientNet-B0) achieved a validation AUC of ~0.95 and accurately identified the correct hemorrhage subtype in roughly 80% of cases. This performance is competitive with other state-of-the-art models in the literature, especially considering our model's relative simplicity and focus on single-slice analysis.

We also showed the effect of excluding the rare epidural hemorrhage class. Removing this class (Model 3) led to a slight increase in the consistency of classification for the remaining subtypes (as evidenced by a higher macro F1-score), at the cost of failing to detect epidural hemorrhages altogether. This underscores the challenge that rare critical findings pose to AI models – strategies beyond outright exclusion, such as separate targeted models or data augmentation, are needed to handle them.

Our findings carry important implications for clinical deployment. A CNN system as presented could serve as a triage tool, alerting radiologists to the presence and type of hemorrhage within seconds of image acquisition. For example, an automated alert for "suspected intraparenchymal hemorrhage" could prioritize that scan for immediate review. Moreover, subtype classification by AI could assist less experienced readers in correctly identifying hemorrhage patterns (for instance, distinguishing subdural vs. epidural), potentially reducing diagnostic errors. The high AUCs achieved indicate that the model is generally reliable in ranking the correct subtype highly; with appropriate threshold tuning, sensitivity can be adjusted to ensure few misses of life-threatening hemorrhages.

However, caution is warranted. The models should be thoroughly validated on diverse patient populations and scanner types. They should ideally be incorporated into a workflow where they augment radiologist decision-making rather than replace it. For instance, an AI suggestion of "subarachnoid hemorrhage" can prompt the radiologist to scrutinize sulcal regions carefully. The relatively lower performance on epidural hemorrhages in our 5-class model highlights that the AI might not catch every case – thus, a "second reader" paradigm, where the AI acts as a safety net or prioritization mechanism, is most appropriate at this stage.

In conclusion, our research contributes a comparative analysis of CNN architectures for ICH subtype classification and demonstrates that even lean 2D CNN models can significantly aid in identifying different types of brain bleeds on CT. With further improvements (handling class imbalance, adding 3D context) and extensive validation, such models have the potential to be integrated into radiology workflows, improving the speed and accuracy of ICH diagnosis. Ultimately, this can lead to faster treatment decisions and better outcomes for patients with acute brain hemorrhages.

7. REFERENCES

1. **Flanders, A. E., Prevedello, L. M., Shih, G., Halabi, S. S., Kalpathy-Cramer, J., Ball, R., et al. (2020).** Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence, 2*(3), e190211. DOI: 10.1148/ryai.2020190211. pubmed.ncbi.nlm.nih.govpubs.rsna.org
2. **Wang, X., Shen, T., Yang, S., Lan, J., Xu, Y., Wang, M., et al. (2021).** A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *NeuroImage: Clinical, 32*, 102785. DOI: 10.1016/j.nicl.2021.102785. (RSNA 2019 Challenge 1st place solution – achieved subtype AUCs ~0.99) pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov
3. **Rajagopal, M., Buradagunta, S., Almeshari, M., Alzamil, Y., Ramalingam, R., & Ravi, V. (2023).** An Efficient Framework to Detect Intracranial Hemorrhage Using Hybrid Deep Neural Networks. *Brain Sciences, 13*(3), 400. DOI: 10.3390/brainsci13030400. (Hybrid CNN-LSTM model on RSNA data; reported ~95% accuracy, F1 ~0.94) mdpi.com
4. **Wu, Y., Supanich, M. P., & Deng, J. (2021).** Ensembled deep neural network for intracranial hemorrhage detection and subtype classification on noncontrast CT images. *Journal of Artificial Intelligence in Medical Sciences, 2*(1), 12–20. DOI: 10.2991/jaims.d.210618.001. (EfficientNet-B0 ensemble; 95.7% accuracy for hemorrhage vs none, strong generalization to external data) atlantis-press.comatlantis-press.com
5. **Ye, H., Gao, F., Yin, Y., Guo, D., Zhao, P., Lu, Y., et al. (2019).** Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European Radiology, 29*(11), 6191–6201. DOI: 10.1007/s00330-019-06163-2. (3D CNN + RNN approach; achieved AUC > 0.8 for all subtypes) pmc.ncbi.nlm.nih.gov
6. **Danilov, G., Kotik, K., Negreeva, A., Tsukanova, T., Shifrin, M., Zakharova, N., et al. (2020).** Classification of Intracranial Hemorrhage Subtypes Using Deep Learning on CT scans. *Studies in Health Technology and Informatics, 272*, 370–373. DOI: 10.3233/SHTI200572. (ResNeXt model tested on Burdenko institute data; >81% accuracy for each subtype without re-training) pubmed.ncbi.nlm.nih.gov
7. **Umapathy, S., Murugappan, M., Thakur, M., & et al. (2023).** Automated computer-aided detection and classification of intracranial hemorrhage using ensemble deep learning techniques. *Diagnostics, 13*(18), 2987. DOI: 10.3390/diagnostics13182987. (Ensemble of SE-ResNeXt and LSTM on RSNA dataset; reported 99.7% accuracy, F1 0.97; noted imbalanced performance with lower sensitivity for EDH)pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov
8. **Burduja, M., Ionescu, R. T., & Verga, N. (2020).** Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans

with convolutional and long short-term memory neural networks. *Sensors, 20*(19), 5611. DOI: 10.3390/s20195611. (3D CNN with bidirectional LSTM on RSNA data; high slice-level AUC ~0.98; EDH sensitivity was lowest ~44%)mdpi.commdpi.com

9. **Lee, J. Y., Kim, J. S., Kim, T. Y., & Kim, Y. S. (2020).** Detection and classification of intracranial hemorrhage on CT images using a novel deep learning algorithm. *Scientific Reports, 10*(1), 20546. DOI: 10.1038/s41598-020-77441-z. (Proposed a custom CNN; one of early works on ICH subtype classification with decent accuracy on a private dataset.)
10. **Korra, S., Mamidi, R., Soora, N. R., Kumar, K. V., & Kumar, N. C. S. (2022).** Intracranial hemorrhage subtype classification using learned fully connected separable convolutional network. *Concurrency and Computation: Practice and Experience, 34*(21), e7218. DOI: 10.1002/cpe.7218. (Demonstrated a lightweight CNN architecture for hemorrhage classification on RSNA data, with competitive performance.)
11. **Hussain, A., Yaseen, M. U., Imran, M., Waqar, M., Akhunzada, A., & Al-Jaafreh, M. (2022).** An attention-based ResNet architecture for acute hemorrhage detection and classification: Toward a Health 4.0 digital twin study. *IEEE Access, 10*, 126712–126727. DOI: 10.1109/ACCESS.2022.3225671. (Used attention mechanisms in a ResNet for hemorrhage subtype classification; achieved high accuracy ~96% on RSNA data.)
12. **Cortés-Ferre, L., Gutiérrez-Naranjo, M. Á., Egea-Guerrero, J. J., Pérez-Sánchez, S., & Balcerzyk, M. (2023).**Deep Learning Applied to Intracranial Hemorrhage Detection. *Journal of Imaging, 9*(2), 37. DOI: 10.3390/jimaging9020037. (Recent work applying various deep learning models to ICH detection; discusses performance and model optimization on RSNA dataset.)
13. **Venugopal, D., Jayasankar, T., Sikkandar, M. Y., Waly, M. I., Pustokhina, I. V., & Pustokhin, D. A. (2021).** A Novel Deep Neural Network for Intracranial Hemorrhage Detection and Classification. *Computational Materials Continuum, 68*(3), 2877–2893. DOI: 10.32604/cmc.2021.015996. (Presented a custom DNN architecture; validated on RSNA challenge data with performance in mid-90% range for detection.)
14. **Tharek, A., Muda, A. S., Hadi, A. B., & Hudin, N. S. (2022).** Intracranial hemorrhage detection in CT scan using deep learning. *Asian Journal of Medical Technology, 2*(1), 1–18. DOI: 10.32896/ajmedtech.v2n1.1-18. (Showed 95% accuracy on a smaller dataset for binary ICH detection using a feed-forward CNN.)
15. **RSNA (2019).** RSNA Intracranial Hemorrhage Detection Challenge (2019) – Dataset. *Radiological Society of North America*. (Description of the RSNA 2019 Brain CT Hemorrhage dataset, containing labeled CT scans for five hemorrhage subtypes)pubs.rsna.org pubmed.ncbi.nlm.nih.gov.
16. Kang, Dong-Wan; Park, Gi-Hun; Ryu, Wi-Sun; Schellingerhout, Dawid; Kim, Museong; Kim, Yong Soo; Park, Chan-Young; Lee, Keon-Joo; Han, Moon-Ku; Jeong, Han-Gil; and Kim, Dong-Eog, "Strengthening Deep- Learning Models for Intracranial Hemorrhage Detection: Strongly Annotated Computed Tomography Images and Model Ensembles" (2023). F aculty , Staff and

Student Publications. 1972. https://digitalcommons.library.tmc.edu/uthgsbs_docs/1972