EfficientNet Sequence Model Pipeline for DICOM-Based Intracranial Hemorrhage Detection with Interpretable Heatmaps

Gabriel Forsberg
(Dated: June 24, 2025)

Intracranial hemorrhage (ICH) remains a time-critical emergency in which rapid, accurate interpretation of non-contrast head CT is essential. While heavyweight convolutional networks can achieve state-of-the-art performance, their computational cost can be restrictive. We therefore designed and compared three lightweight, interpretable pipelines that share an EfficientNet-B0 backbone but incorporate progressively richer contextual cues: (i) a plain 2-D slice classifier, (ii) a "2.5-D" variant that concatenates neighboring slices, and (iii) a sequence model that fuses slice-to-slice dynamics through a single-layer LSTM. All models were trained on 50,863 slices from 1,000 studies in the RSNA ICH dataset, windowed into bone, subdural and brain settings. The plain 2-D benchmark achieved a mean F1-score of 0.77 across the five RSNA hemorrhage subtypes. Simply adding adjacent slices (2.5-D) did not improve performance (mean F1 = 0.76), whereas explicit temporal modeling with the LSTM raised the mean F1 to 0.79 and delivered the best ROC curves (micro-AUC = 0.98). Gradient-weighted class-activation maps revealed that the LSTM consistently focused on anatomically plausible bleed regions, providing transparent visual rationales.

Teaching Assistant: Yu-Wei Chang

I. INTRODUCTION

Intracranial hemorrhage (ICH) refers to bleeding within the skull, either confined to the brain parenchyma or occurring in the spaces between the brain and its protective membranes [1]. This life-threatening condition demands prompt medical intervention to minimize the risk of permanent neurological deficits or fatality. A recent global burden study estimated that approximately 3.4 million individuals experienced ICH in 2019, leading to an estimated 2.8 million deaths worldwide [2].

In a 2003 investigation evaluating the performance of board-certified radiologists in interpreting emergency head CT scans via teleradiology, 716 emergency head CTs were reviewed [3]. The authors reported a sensitivity of 85% and a specificity of 99.8% for identifying any form of intracranial hemorrhage.

More recently, deep learning approaches have been employed to enhance both the speed and accuracy of ICH detection. In their 2025 meta-analysis, Karamian and Seifi synthesized results from multiple studies of deep learning algorithms applied to non-contrast CT (NCCT) and found a pooled sensitivity of 0.92, specificity of 0.94, positive predictive value (PVP) of 0.84, negative predictive value (NPV) of 0.97, and an area under the curve (ROC) of 0.96 [4].

In this report, three lightweight and interpretable models, building on the EfficientNet architecture with increasing complexity, are presented and compared. They were designed to strike a balance between complexity, computational efficiency and accuracy. Combined with Grad-CAM heatmaps, highlighting regions of interest within each CT slice, the proposed frameworks aims to achieve achieve robust hemorrhage detection while also providing clear visual explanations that can support clinical decision-making. We evaluate the models on a diverse, non-contrast CT dataset (RSNA), demonstrating that they maintain high accuracy while reducing training and inference time compared to heavier architectures. Finally, we discuss their respective benefits and drawbacks as well as how they could be integrated into existing radiology workflows to facilitate rapid triage and bolster physician confidence in AI-assisted diagnosis.

II. METHOD

All models presented in this report use the EfficientNet-B0 image classification model as the backbone, developed by researchers at google in 2019 [5]. The model utilizes mobile inverted bottlenecks as well as squeeze-and-excitation optimization to achieve up to 10 times better efficiency than comparable models [5, 6], which makes it suitable for computationally constrained environments while preserving high accuracy.

A total of 50,863 CT images (N=1,000 patients) were randomly drawn from the RSNA Intracranial Hemorrhage Detection challenge dataset and partitioned into training, validation and test (70/15/15). See Table I for distribution of hemorrhages present in the subset.

Prior to training, every CT study was also re-windowed into three standard Hounsfield-unit ranges—[40, 80] HU for bone, [40, 200] HU for subdural blood and [500, 2000] HU for brain parenchyma—to enhance contrast of the structures of interest. From each re-windowed volume we extracted axial slices and linearly normalized pixel values to [0, 1]. These three gray-scale windows were then concatenated into a single three-channel image, providing the model with complementary tissue contrasts in one input tensor.

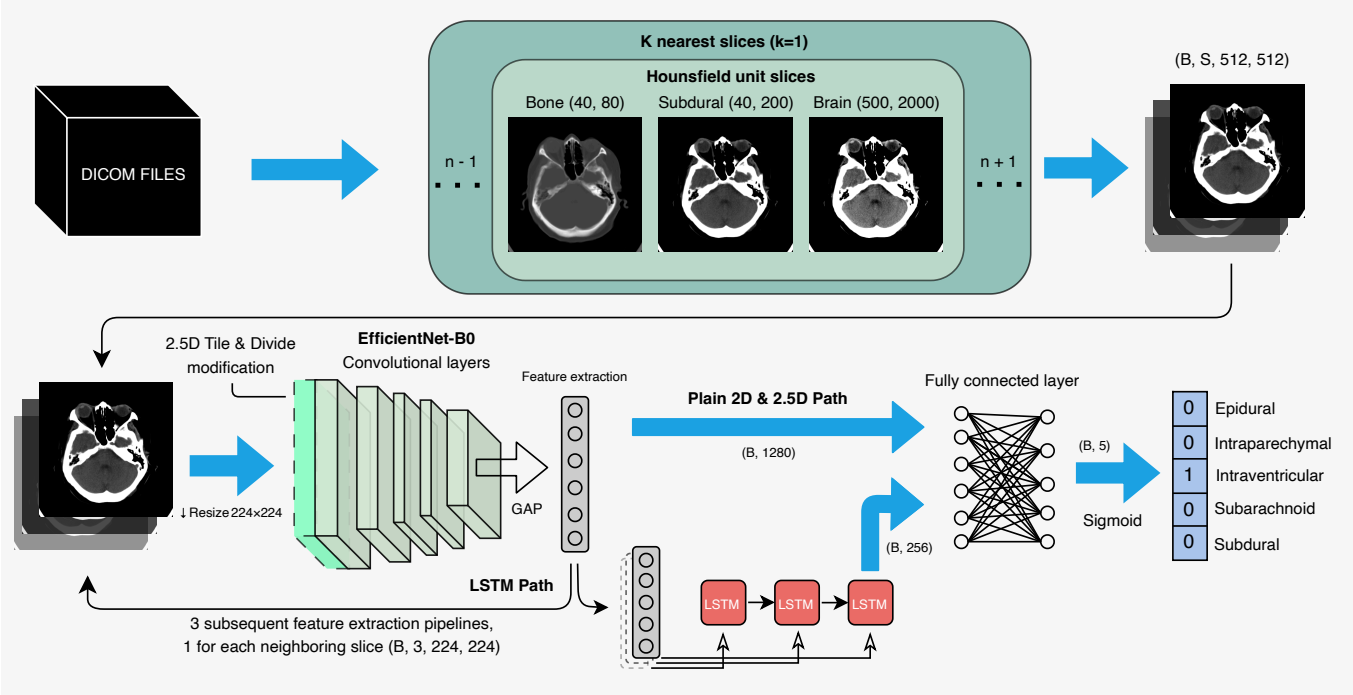

FIG. 1. **End-to-end inference pipeline for intracranial-hemorrhage detection.** All three branches share identical preprocessing: each CT slice is windowed into Bone, Subdural and Brain settings and stacked. **1. Plain 2-D** A single three-channel slice $(B, 3, 224, 224)$ is fed to EfficientNet-B0; the pooled $(B, 1280)$ feature vector maps directly to five logits through a single fully connected layer. **2. 2.5-D** The central slice plus two neighbors (nine channels) form $(B, 9, 224, 224)$; the stem is widened 3→9 via "Tile & Divide", after which the pathway mirrors branch 1. **3. LSTM sequence** Window triplets are encoded slice-by-slice, yielding a sequence $(3, B, 1280)$; a 256-unit LSTM fuses context and a linear head outputs the logits. In every branch, sigmoid activation converts the logits to probabilities for the five RSNA haemorrhage subtypes.

TABLE I. **Slice counts by hemorrhage type**

Hemorrhage Type	# Slices
Epidural	240
Intraparenchymal	2 853
Intraventricular	1 962
Subarachnoid	2 627
Subdural	3 180

All three models follows the structure of the chart in Figure 1, with slight variations in implementation and complexity. The first model, *Plain 2D*, serves as a benchmark for the subsequent models and employs the EfficientNet-B0 model with the aforementioned Hounsfield windows to make slice-level predictions of the type hemorrhage present or absent in each slice. First, each window stack is fed through the EfficientNet backbone, which after a global-average pooling layer, results in feature vectors of size (1, 1280). Each batch (B) of such feature vectors are then put trough a linear fully connected layer with a sigmoid activation function to produce the final per-class probabilities.

The second model, *2.5D Feature extraction*, expands the context available to the network by also providing it with information from its two neighboring slices. This is done by replacing the stem of the network to accept 9 input channels instead of the original 3-channels. The weights of the first convolutional layer is simply copied three times and then concatenated, "*tiled*", along the channel axis. To avoid potentially blowing up early activations and preserve the expected variance of the output, the weights are scaled back down. This "*Tile and Divide*" modification allows the network to gain some spatial awareness while the overall magnitude of the convolution outputs stays consistent with the pretrained model.

The third model, *Sequence branch with LSTM*, adds further complexity by introducing a linear sequence LSTM model to allow it to learn explicit slice-to-slice dynamics, so far not present in the previous models. Instead of feeding the full stack of 9 slices (3×3 windows), each of the 3 window triplets (center slice and its two neighbors) are subsequently fed through the network, producing three separate feature vectors. These slice-level feature vectors are then fed, in scan order, to a single-layer LSTM (hidden_dim = 256). The final hidden state of size $(B, 256)$ is then mapped to the final 5 predictions by a lightweight fully-connected layer.

After training, we sweep class-specific decision thresholds on the validation set to maximise F1. The result-

ing optimal thresholds are then frozen and applied unchanged to the held-out test set.

In addition to the quantitative metrics reported for each architecture, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to gain qualitative insight into the spatial regions that most strongly influenced every model's predictions. The gradients of the predicted class scores with respect to the final convolutional feature map was back-propagated, pooled, and projected onto the input image, yielding a coarse heatmap that highlights regions deemed salient by the network. To make the resulting saliency maps easier to interpret, the heat-maps were (i) normalized to the full 0–1 range, (ii) up-sampled to the original 512×512 resolution with bilinear interpolation, and (iii) overlaid in false-color on the corresponding CT slice.

III. RESULTS AND DISCUSSION

The resulting per-class F1 scores on the test set of each model and their respective micro-averages is shown in Table II. Higher numbers indicate better bleed detection.

TABLE II. **F1 Scores for different hemorrhage types**

Hemorrhage	2D	2.5D	LSTM
Epidural	0.75	0.75	0.77
Intraparenchymal	0.79	0.78	0.82
Intraventricular	0.77	0.76	0.75
Subarachnoid	0.75	0.73	0.78
Subdural	0.79	0.80	0.84
Average	0.77	0.76	0.79

As can be seen for the table, there is no discernible improvement in the 2.5D model, that added the neighboring slices to the convolutional network, over the standard 2D EfficientNetB-0 model. In fact, it slightly reduced the F1 scores on average. The sequence LSTM model on the other hand, showed an improvement in 4 of 5 of the bleed types, especially subdural (+0.05, +0.04), showing that adding slice-to-slice context helps the network in general make more accurate calls than the 2-D or 2.5-D versions.

To further analyze how well each model distinguish between the different bleed types, regardless of the set threshold from the grid search on the validation set, we can compare the models Receiver-operating characteristic curves (ROC), see Figure 2.

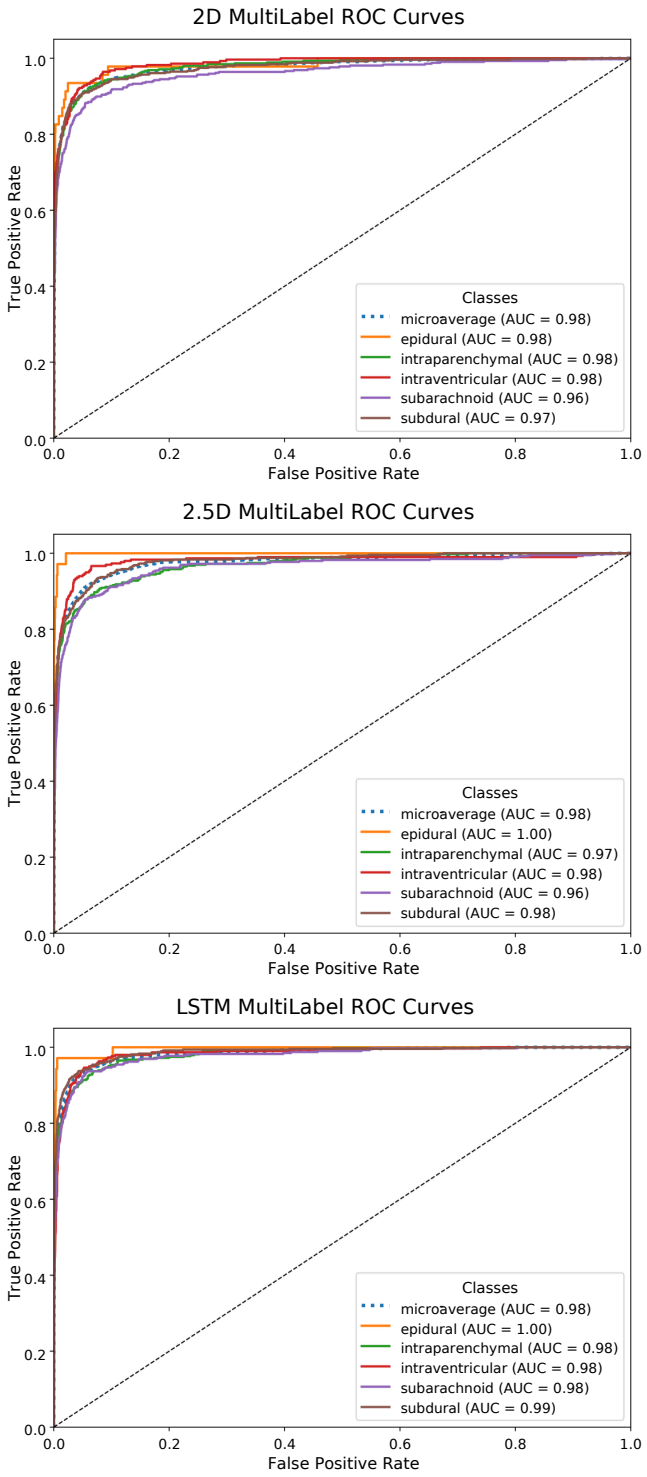

FIG. 2. **Receiver-operating characteristic (ROC) curves for multi-label intracranial hemorrhage classification on the test set.** The blue dotted line shows the micro-average ROC curve (AUC = 0.98), summarizing overall slice-level discrimination across all hemorrhage types, while the solid curves represent each subtype's ROC: Epidural, Intraparenchymal, Intraventricular, Subarachnoid, and Subdural. The black dashed diagonal denotes chance performance (AUC = 0.50).

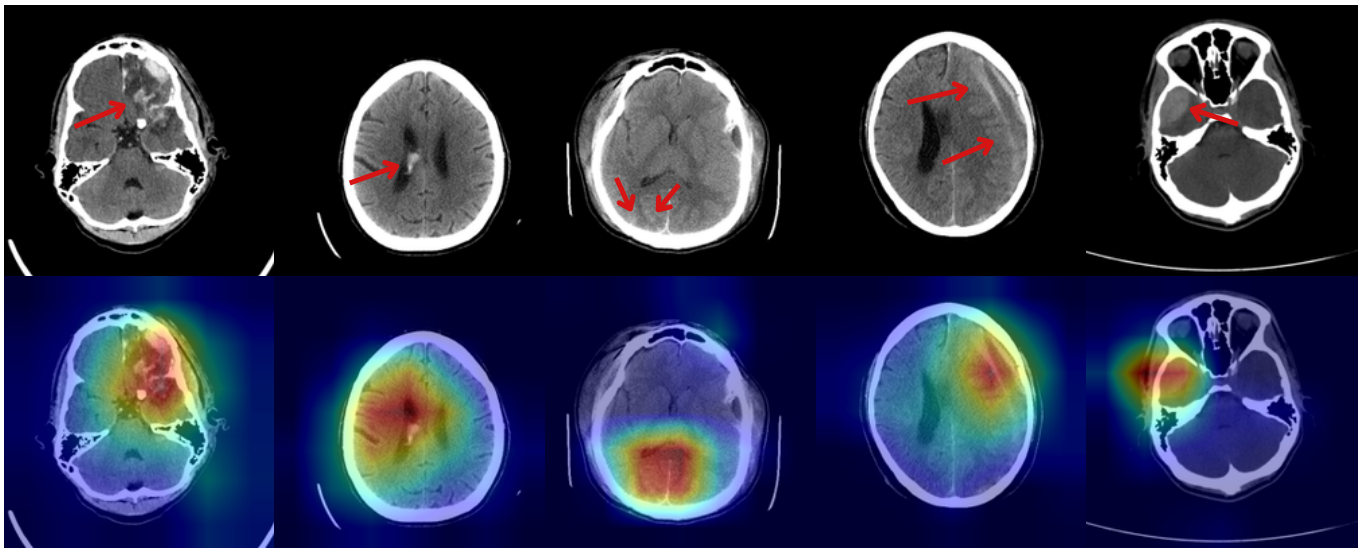

FIG. 3. **CT images and corresponding Grad-CAM saliency maps for five intracranial hemorrhage types.** Top row: Axial non–contrast CT slices demonstrating (from left to right) (A) intraparenchymal hemorrhage, (B) intraventricular hemorrhage, (C) subarachnoid hemorrhage, (D) subdural hemorrhage, and (E) epidural hemorrhage. Bottom row: Grad-CAM overlays computed from the sequence LSTM model, projected onto the same CT slices. Warmer colors indicate regions of greatest model attention when classifying each hemorrhage subtype.

From the figure, we can see that the 2.5D network (middle graph) actually scores higher on two of the hemorrhage classes compared to the plain 2D case (top graph), showing that the extra context helped the classifier rank positives ahead of negatives more reliably in some cases. However, these ranking gains did not translate into the threshold-dependent F1 metric, likely because the optimal cut-off chosen on the validation grid shifted unfavorably once applied to the test set.

For the sequence LSTM model (bottom graph), all ROC curves performed better or equally as good as the two other architectures, again showing the performance increase of the added slice-to-slice dynamics.

In Figure 3, one example of each hemorrhage type and their respective Grad-CAM saliency maps are shown, extracted from the best performing sequence LSTM model. The selected images show that the gradients of the predicted classes with respect to the final convolutional layer, can correctly highlight regions of the CT scans where a bleed is present, and thus increase the exlainability of the final predictions. However, it is not guaranteed that the resulting saliency maps directly corresponds to bleed regions in the anatomical sense. Grad-CAM highlights where the network's last convolutional feature maps change the logit of a given class the most, and that influence may arise from subtle contextual cues (e.g. midline shift, ventricle shape) rather than from voxels that actually contain hemorrhage. Consequently, saliency maps should be interpreted as suggestive rather than diagnostic. In clinical deployment they are best used to draw the reader's attention to candidate regions while preserving the radiologist's final responsibility for verification.

IV. CONCLUSIONS AND OUTLOOK

Our experiments show that progressively adding context to an EfficientNet-B0 backbone can raise slice-level intracranial-hemorrhage performance while keeping the model architecture lean and efficient to run on consumer hardware, including directly on CPU. The plain 2-D model already reached an average F1 of 0.77, but explicitly modeling slice-to-slice continuity with a lightweight LSTM lifted the score to 0.79 and produced the strongest ROC curves across all five hemorrhage subtypes, while Grad-CAM heat-maps confirmed that the network attends to anatomically plausible regions. Simply stacking neighboring slices in a "2.5-D" input was less helpful, improving rank-ordering for some classes yet failing to boost threshold-dependent F1, suggesting that context must explicitly be learned rather than hinted.

Future work might examine supplying the LSTM with a wider window of neighboring slices or integrating slices from multiple imaging planes to better capture and differentiate three-dimensional features. Methods of improving the consistency of the Grad-CAM explanations across adjacent slices, such as enforcing temporal-smoothness constraints or regularizing the maps with weak anatomical priors, could also potentially further boost the detector's clinical reliability.

V. DATA AND CODE AVAILABILITY

The RSNA dataset is available for download at: www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/data

All code used for this report is available at: github.com/TheGabbe/RSNA-EfficientNetModels

[1] J. A. Caceres and J. N. Goldstein, Intracranial hemorrhage, Emergency Medicine Clinics of North America **30**, 771 (2012), used as a source to define intracranial hemorrhage.

[2] T. Sun, Y. Yuan, K. Wu, Y. Zhou, C. You, and J. Guan, Trends and patterns in the global burden of intracerebral hemorrhage: a comprehensive analysis from 1990 to 2019, Frontiers in Neurology **14**, 1241158 (2023), provides relevant statistics for the global burden of ICH, highlighting a great need for further research.

[3] W. K. Erly, B. C. Ashdown, R. W. Lucio, R. F. Carmody, J. F. Seeger, and J. N. Alcala, Evaluation of emergency ct scans of the head: is there a community standard?, American Journal of Roentgenology **180**, 1727 (2003), relevant because it provides a performance overview of board-certified radiologists that can be used to benchmark machine aided or fully automatic diagnoses.

[4] A. Karamian and A. Seifi, Diagnostic accuracy of deep learning for intracranial hemorrhage detection in non-contrast brain ct scans: A systematic review and meta-analysis, Journal of Clinical Medicine **14**, 2377 (2025), relevant because it provides an overview of SOTA deep learning approaches and their pooled scored perfomances, which can be used to evaluate the performance of the models presented in this report.

[5] M. Tan and Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks (2020), relevant because it itroduces the backbone model used in all three frameworks., arXiv:1905.11946 [cs.LG].

[6] M. Tan and Q. V. Le, Efficientnet: Improving accuracy and efficiency through automl and model scaling (2019), google AI Blog, accessed 2025-05-09, relevant because it provides additional details about the EfficientNet architecture.