**A Comparative Evaluation of Baseline and Deep Learning Models for Intracranial Hemorrhage Detection**

Owen Hughes

University of Oregon

[obh@uoregon.edu](mailto:obh@uoregon.edu)

***Abstract***

I evaluate three models for intracranial hemorrhage detection on RSNA 2019 head CTs. A 20 000-slice subset (50 % positive) was split patient-wise 60 / 20 / 20, and each slice converted into a 3-channel 160 × 160 PNG using brain, subdural, and bone windows. Models include (1) logistic regression on flattened images (AUC ≈ 0.67, F1 ≈ 0.58), (2) a SmallResNet trained from scratch with AdamW + OneCycleLR (AUC = 0.8575, F1 = 0.7840), and (3) a ResNet-18 fine-tuned end-to-end (AUC = 0.8813, F1 = 0.7827). Grad-CAMs reveal sharper lesion localization in ResNet-18 versus more diffuse attention in the small network. These results show that multi-window inputs and modern training schedules yield substantial gains over the baseline.

**I. INTRODUCTION**

Intracranial hemorrhage (ICH) is one of the most life-threatening findings on head CT, with mortality rising sharply if not diagnosed and treated within the first hour. Radiologists must rapidly scan hundreds of axial slices to identify bleeds—an arduous task prone to human error under time pressure. Automated methods for ICH detection promise to reduce turnaround times and improve patient outcomes by flagging critical studies immediately upon acquisition.

Motivated by the RSNA 2019 Intracranial Hemorrhage Detection challenge, numerous top‐ranked solutions have relied on large ensembles of 2-D, 2.5-D, and 3-D networks

with extensive compute budgets, achieving exceptional accuracy but at the cost of interpretability and real-world deployability. In this work, I explore a streamlined three-model pipeline that balances simplicity and performance: a logistic-regression baseline on down-sampled pixels, a compact residual CNN trained from scratch, and an ImageNet-pre-trained ResNet-18 fine-tuned end-to-end.

To enable clinically meaningful feature learning, each CT slice is windowed into three channels (brain, subdural, bone), resized to 160 × 160, and used to train on a patient-stratified 60 / 20 / 20 split of 20 000 slices. I employ modern training techniques—AdamW optimization with OneCycleLR scheduling—and validate on a held-out test set. Grad-CAM visualizations further reveal each model’s attention patterns. The remainder of this report details the dataset, methods, experiments, and results, demonstrating that a focused combination of multi-window preprocessing and transfer learning yields near state-of-the-art AUC with a lightweight, reproducible pipeline.

**II. BACKGROUND**

Intracranial hemorrhage (ICH) detection on non‐contrast head CT is a critical clinical task that demands rapid and accurate decision-making. CT images inherently span a wide range of tissue densities—from cerebrospinal fluid to bone—so radiologists routinely adjust window‐level (WL) and window‐width (WW) settings to highlight blood, soft tissue, or bone structures. Historically, rule‐based and classical machine learning methods (e.g., support‐vector machines on hand‐crafted features) demonstrated moderate success but required extensive feature engineering and lacked robustness across scanners and patient populations.

In recent years, convolutional neural networks (CNNs) have revolutionized medical image analysis by learning hierarchical representations directly from pixel data. Early work applied 2-D CNNs on single‐window slices, while subsequent studies developed 2.5-D models that incorporate adjacent slices or 3-D CNNs to leverage volumetric context. Transfer learning—fine-tuning ImageNet‐pretrained networks on CT data—has proven especially effective, often accelerating convergence and improving accuracy when labeled medical volumes are scarce.

Grad‐CAM (gradient‐weighted class activation mapping) has become a standard tool for visualizing CNN decision‐making: it highlights the regions in an image that most strongly influence a model’s output. By combining multi‐window inputs with compact CNN architectures and transfer-learning paradigms, my work builds on these advances to deliver a lightweight yet high‐performance pipeline suitable for rapid ICH triage in clinical settings.

**III. METHODS**

**1. Preprocessing and Data Preparation**  
I first sampled a balanced subset of 20 000 axial CT slices (10 000 positive, 10 000 negative) from the full RSNA training set, stratified by patient to avoid data leakage. The master sample\_20k.csv was split patient-wise into 60 % train, 20 % dev, and 20 % test. Each DICOM slice was converted to a 3-channel 160 × 160 PNG by applying three window-level settings:

* **Brain window** (WL = 40, WW = 80) for parenchymal contrast
* **Subdural window** (WL = 75, WW = 215) for extra-axial blood
* **Bone window** (WL = 600, WW = 2000) for high-contrast structures

These channels were stacked as a pseudo-RGB image, resized with area interpolation, and written to data/processed/160. Training‐time augmentations included random horizontal flips and rotations up to ±10°.

**2. Models**  
Three architectures were evaluated:

* **Logistic Regression (Baseline)** Each PNG was loaded as a single‐channel 160² grayscale image, flattened to a 25 600-dim vector, and fit with L2-regularized logistic regression (class\_weight='balanced').
* **SmallResNet (Scratch)** A custom residual CNN with three residual blocks (channel widths 32→64→128→256), each incorporating a squeeze-and-excitation unit and 3×3 convolutions, followed by global average pooling and a single-unit output layer. Input channels = 3. Total parameters ≈ 1.2 M.
* **ResNet-18 Fine-Tune** The standard ImageNet-pretrained ResNet-18, modified to accept 3-channel 160×160 inputs (conv1 kernel updated) and a single-unit output layer. All layers were fine-tuned.

**3. Training and Hyperparameter Tuning**  
Models were trained with **binary cross-entropy with logits** loss on positive (any=1) vs. negative slices. The optimizer was **AdamW** (weight\_decay=1e-4) combined with **OneCycleLR** (peak LR = chosen LR; total steps = epochs×(N\_train/B)). Early stopping with patience = 3 epochs prevented over-training.

* **Hyperparameter sweep**: learning rates {1e-3, 3e-4} for SmallResNet and {1e-4, 3e-5} for ResNet-18; batch sizes 32 and 16 respectively; each trial ran 6 epochs on train/dev only. The best dev AUC determined the LR.
* **Final training**: selected LR on the combined train+dev set for up to 15 epochs, with early stopping. All training ran on Apple-Silicon MPS unless unavailable.

Data loading and batching used PyTorch DataLoader with 4 workers and normalized inputs. The entire pipeline—from DICOM windows to PNG conversion, through model training and Grad-CAM visualization—was fully automated via CLI scripts (preprocess.py, train.py, tune.py, evaluate.py, cam.py), ensuring reproducibility.

**IV. EXPERIMENTS**

**1. Hyperparameter Tuning (Dev Set)**  
I tuned learning‐rate on the dev split (20 % of data) for the two CNN models, holding batch size fixed.

**SmallResNet**

| **Learning rate** | **Batch** | **Dev AUC** |
| --- | --- | --- |
| 1 × 10⁻³ | 32 | 0.7597 |
| 3 × 10⁻⁴ | 32 | 0.7492 |

The best dev AUC was **0.7597** at **LR = 1 × 10⁻³**, which was used for final training.

**2. Final Test‐Set Metrics**  
Each model was retrained on the combined train+dev set (80 % of data) using its selected hyper‐parameters, then evaluated once on the held‐out test split (20 %). Early stopping (patience = 3) prevented over‐training.

| **Model** | **Test AUC** | **Test F1** |
| --- | --- | --- |
| Logistic Regression (baseline) | 0.6731† | 0.5762† |
| SmallResNet (3-window) | 0.8575 | 0.7840 |
| ResNet-18 (ImageNet FT) | 0.8813 | 0.7827 |

† *Logistic-regression test metrics were estimated by fitting on train+dev and predicting on the test split.*

**3. Qualitative Analysis (Grad-CAM)**  
To visualize model focus, I generated Grad-CAM overlays on one true‐positive and one false‐negative test slice:

| **Slice ID** | **SmallResNet CAM** | **ResNet-18 CAM** |
| --- | --- | --- |
| **True-positive** (ID f1c8ce782) | A close-up of a scan of a human body  AI-generated content may be incorrect. | A rainbow colored image of a human skull  AI-generated content may be incorrect. |
| **False-negative** (ID ad7f85865) | A blue circle with a bright light  AI-generated content may be incorrect. | A blue circle with a light in the center  AI-generated content may be incorrect. |

* **True-positive:** ResNet-18’s CAM (right) is sharply focused on the hemorrhage region, whereas the SmallResNet CAM (left) spreads activation more diffusely across the cortex.
* **False-negative:** SmallResNet shows a spurious hot‐spot in the ventricle, explaining its misclassification, while ResNet-18 correctly suppresses activation, reflecting its higher specificity.

With these quantitative and qualitative results, the report demonstrates clear performance gains from baseline through custom CNN to transfer‐learning, validating the multi-window preprocessing and modern training strategies.

**V. CONCLUSION**

In this project, I developed and evaluated a streamlined pipeline for automated intracranial hemorrhage detection on head CT scans, using three progressively sophisticated models: a logistic‐regression baseline, a custom 3-window SmallResNet trained from scratch, and an ImageNet-pretrained ResNet-18 fine-tuned end-to-end. By constructing a balanced 20 000-slice dataset and converting each slice into three contrast windows (brain, subdural, bone), I provided each model with richer input than single-window approaches. Modern training techniques—AdamW optimization with OneCycleLR scheduling and early stopping—ensured efficient convergence on Apple-Silicon hardware.

Quantitatively, the SmallResNet achieved a Test AUC of 0.8575 (F1 = 0.7840), a significant jump over the logistic baseline (AUC ≈ 0.673, F1 ≈ 0.576), while ResNet-18 further improved to an AUC of 0.8813 (F1 = 0.7827). Qualitative Grad-CAM visualizations confirmed that ResNet-18 localized hemorrhage regions more sharply, whereas the smaller network exhibited more diffuse or occasionally spurious activations. These results demonstrate that multi-window inputs coupled with transfer learning can yield near-state-of-the-art performance with a concise, reproducible codebase.

**Limitations:**

* The pipeline processes each CT slice independently, ignoring 3-D spatial context that could improve detection of subtle bleeds.
* Labels are at the slice level rather than the voxel level, introducing potential annotation noise.
* I used a relatively small subset (20 000 of ≈ 750 000 slices) for computational tractability; performance may scale further with more data.

**Future Work:**

1. **Incorporate 2.5-D or 3-D context** by stacking adjacent slices or using volumetric architectures to capture continuity of hemorrhages across slices.
2. **Ensemble multiple backbones** (e.g., SmallResNet + EfficientNet + ResNet-18) to reduce variance and boost robustness.
3. **Expand dataset size** for full-scale training and explore semi-supervised or self-supervised pre-training to leverage unlabeled data.
4. **Clinical validation** by testing on external CT datasets and integrating model outputs with PACS systems for real-world deployment.

Overall, this work illustrates that carefully chosen architectural simplicity, enriched input channels, and modern optimization strategies can achieve high accuracy for critical medical imaging tasks within practical compute constraints.

**REFERENCES**

1. RSNA Intracranial Hemorrhage Detection. Kaggle Competition, 2019. Accessed May 2025. Available at: <https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection>
2. He, K., Zhang, X., Ren, S., & Sun, J. “Deep Residual Learning for Image Recognition.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
3. Hu, J., Shen, L., & Sun, G. “Squeeze-and-Excitation Networks.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
4. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. “Grad-CAM: Visual Explanations from Deep Networks via Gradient‐Based Localization.” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
5. Loshchilov, I., & Hutter, F. “Decoupled Weight Decay Regularization.” *International Conference on Learning Representations (ICLR)*, 2019.
6. Smith, L. N. “A Disciplined Approach to Neural Network Hyper‐Parameters: Part 1—Learning Rate, Batch Size, Momentum, and Weight Decay.” *arXiv preprint arXiv:1803.09820*, 2018.
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … & Chintala, S. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.