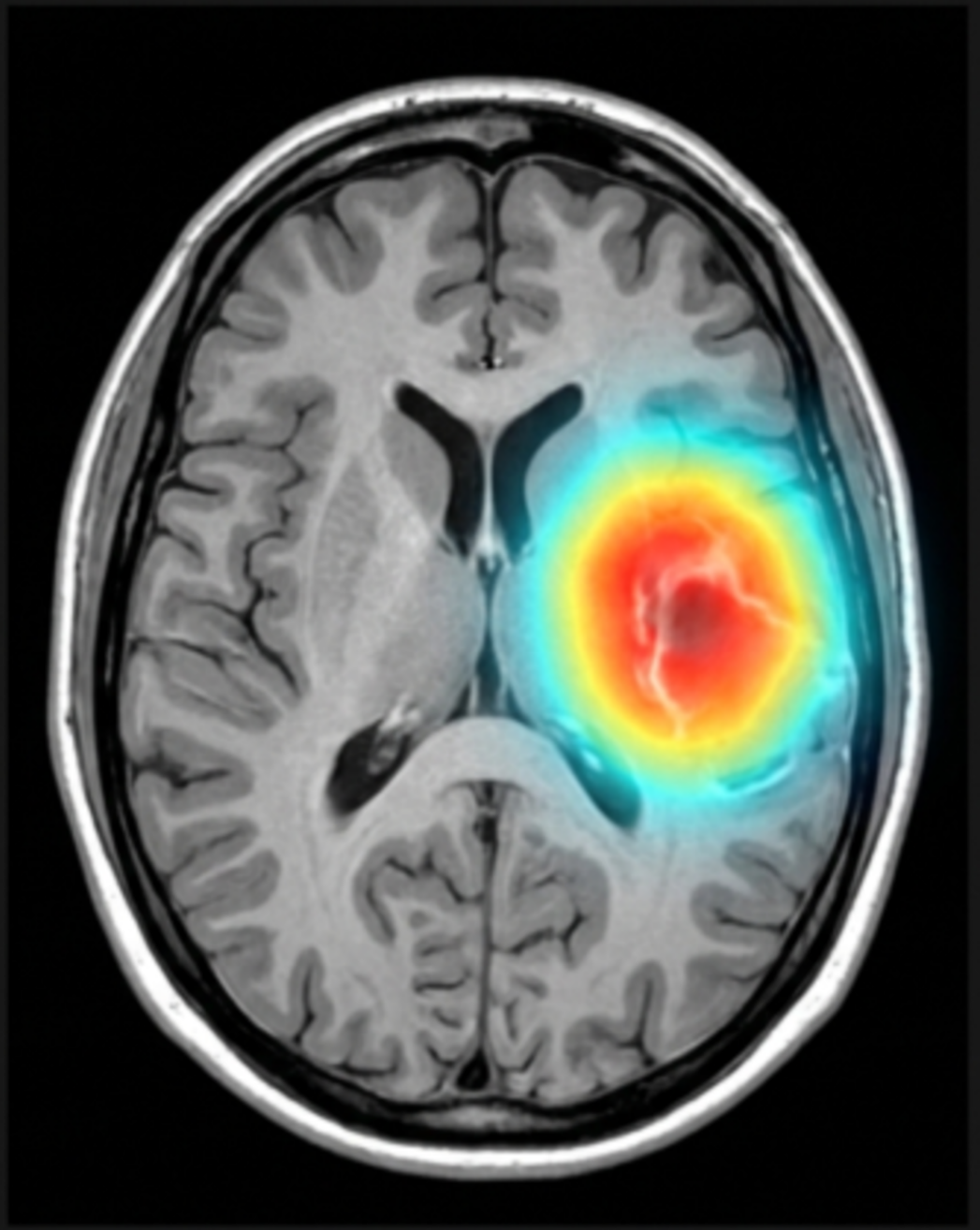
Explainable Deep Learning for Brain Tumor Classification

Visualizing Model Decisions with Grad-CAM from MRI

Author: Waqas, Al Ghazali University (GU Tech), Karachi

Medical Physics Conference for Pakistani Students 2026

Track: AI × Health Science

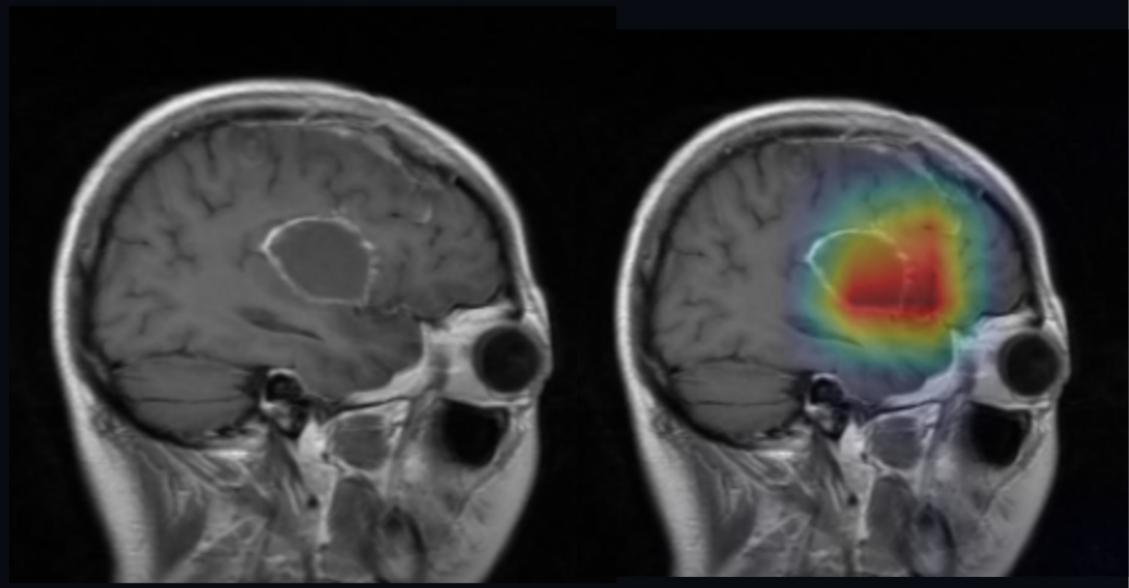


From Computer Science to Medical AI

M. Waqas - Presenting Author 24G-BCS023@agu.edu.pk
M.Anas - Co Author 24G-BCS013@agu.edu.pk
M.Moosa Khan - Co Author 24G-BCS019@agu.edu.pk

We are three fifth-semester computer science students at Ghazali University in Karachi.

Even a correct label can come from image borders or surrounding anatomy. We examined which regions contributed to each prediction.

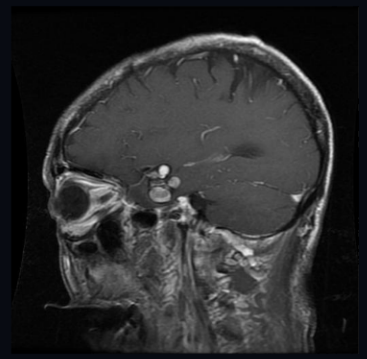


MRI INPUT GRAD-CAM OVERLAY

Can we evaluate both what an AI predicts and where it focuses?

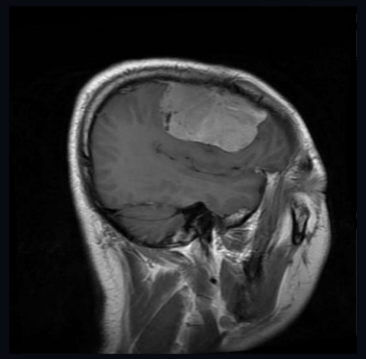
Beyond Accuracy: Toward Explainable Predictions

Compare four architectures on the same four MRI labels and inspect the regions behind each prediction.



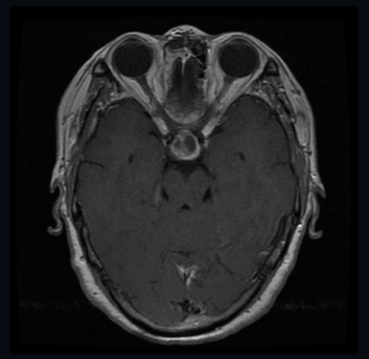
Glioma

Arises from glial cells



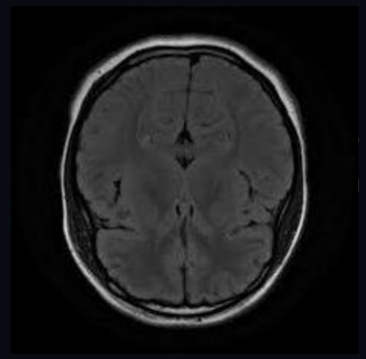
Meningioma

Arises from the meninges



Pituitary tumor

Involves the pituitary gland



No tumor

No labeled tumor

ARCHITECTURES

Custom CNN (from scratch) / ResNet50 (residual connections)
EfficientNet-B0 (efficient convolutions) / DenseNet121 (dense connections)

From MRI Images to Explainable Predictions

Nickparvar's Kaggle compilation contains 7,023 images. The counts below describe the project's working folders.

01 MRI FOLDERS	02 AUDIT	03 SPLIT	04 MODELS	05 TEST	06 EXPLAIN
5,600 train 1,600 test	590 train images matched test images. 337 duplicate pairs within training.	4,673 cleaned 3,972 train 701 validation	One baseline CNN and three ImageNet-pretrained networks.	1,600 held out 1,179 in the leak-free subset.	Grad-CAM maps for correct and incorrect predictions.

The audit found near-duplicates in raw training for 421 of 1,600 test images. The additional subset excludes those 421.

We augmented training images, weighted the loss by class and selected checkpoints using validation macro F1.

Transfer Learning Outperformed the Baseline

HELD-OUT TEST / 1,600 IMAGES / 400 PER CLASS

		MACRO F1	MACRO AUC
Custom CNN	74.25%	0.7396	0.8994
EfficientNet-B0	85.81%	0.8558	0.9666
DenseNet121	91.06%	0.9101	0.9813
ResNet50	91.25%	0.9120	0.9797

ResNet50 led by three correct classifications over DenseNet121, which had the highest macro AUC. This small gap does not establish superiority.

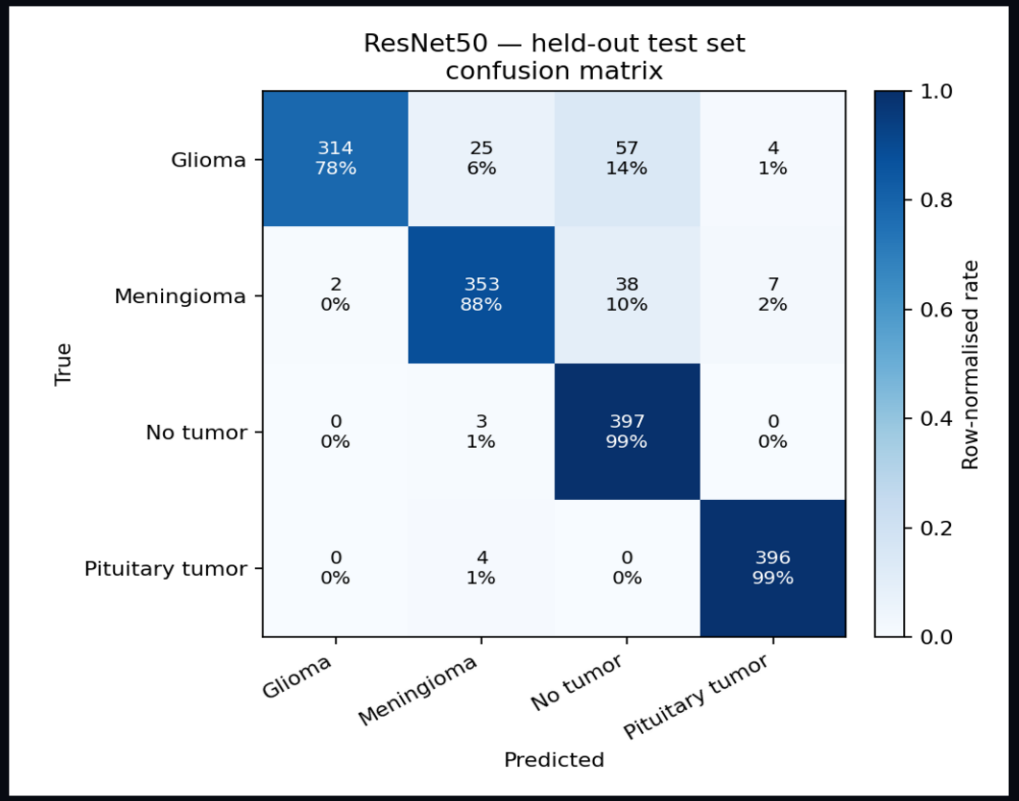
What Happens When We Examine the Mistakes?

LEAK-FREE SUBSET / 1,179 IMAGES

Custom CNN	79.64%
EfficientNet-B0	87.79%
DenseNet121	91.94%
ResNet50	91.52%

Only 89 of 400 no-tumor images remain. The changed class mix limits comparisons with full-test accuracy.

Glioma recall (n = 400) was 0.785 for ResNet50 and 0.788 for DenseNet121. Recall is the fraction of true gliomas identified.



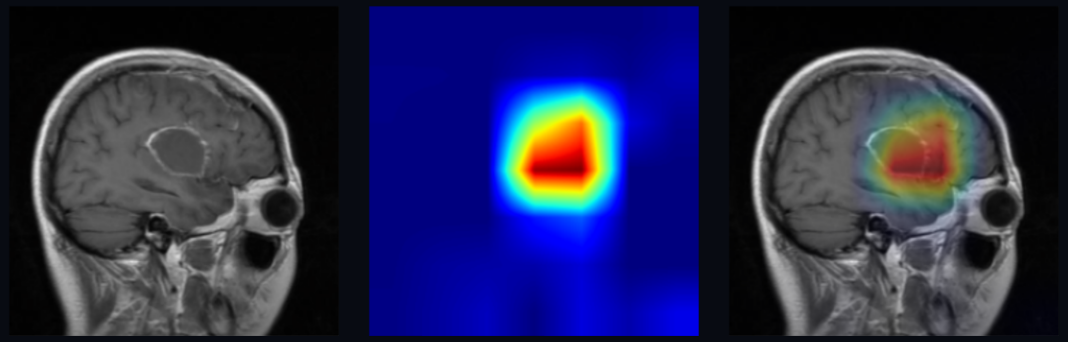
ORIGINAL TEST / RESNET50 / N = 1,600

Tumors called “no tumor” (n = 1,200): ResNet50 7.92%, DenseNet121 8.42%, EfficientNet-B0 12.75%.

Where Did the Model Look?

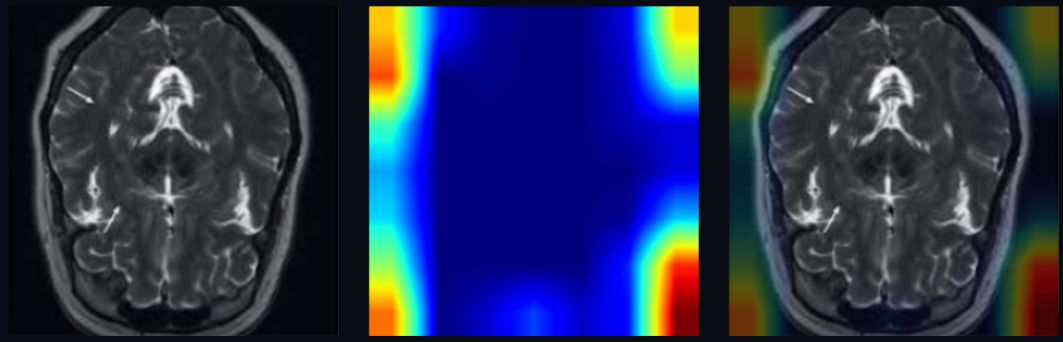
Gradient-weighted Class Activation Mapping (Grad-CAM) weights feature maps by their gradients to highlight regions associated with the predicted class.

CORRECT / GLIOMA CALLED GLIOMA



Original MRI Grad-CAM Overlay

INCORRECT / NO TUMOR CALLED MENINGIOMA



Original MRI Grad-CAM Overlay

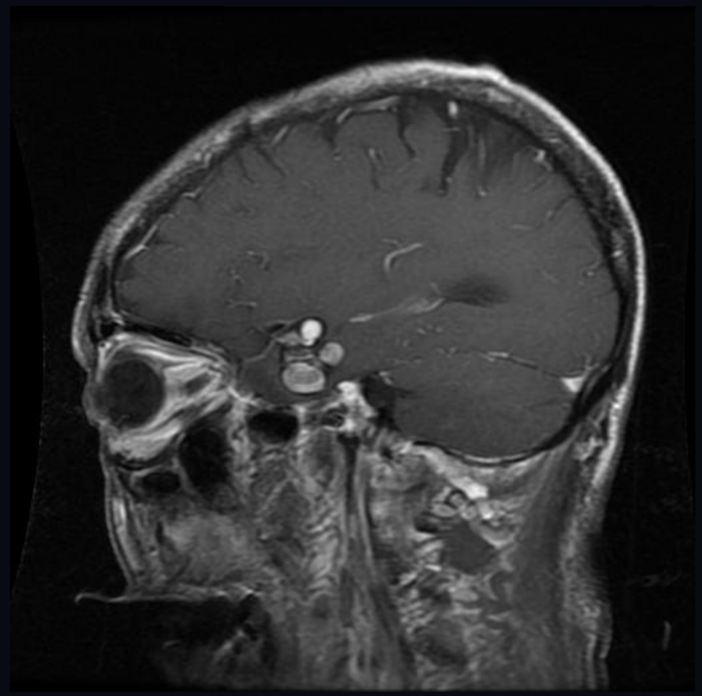
0.89 / 8 correct examples

0.58 / 6 incorrect examples

Mean share of activation in the foreground, measured on selected ResNet50 examples. This coarse measure cannot establish causality or validate a tumor boundary.

Accurate Predictions Are Only Part of the Story

Pretrained models reached about 91% accuracy on 1,600 held-out slices. Duplicate screening changed the test mix, and Grad-CAM revealed some decisions linked to peripheral anatomy.



LIMITATIONS AND NEXT STEPS

This prototype uses 2D slices from limited sources and lacks external clinical validation. Next steps include patient-level splits, independent multicenter testing, stronger robustness checks and comparison with expert tumor annotations.

“The goal is not only to build AI that predicts accurately, but to understand when and why its predictions deserve scrutiny.”

View the project on GitHub ↗

M. Waqas · M. Anas · Moosa Khan