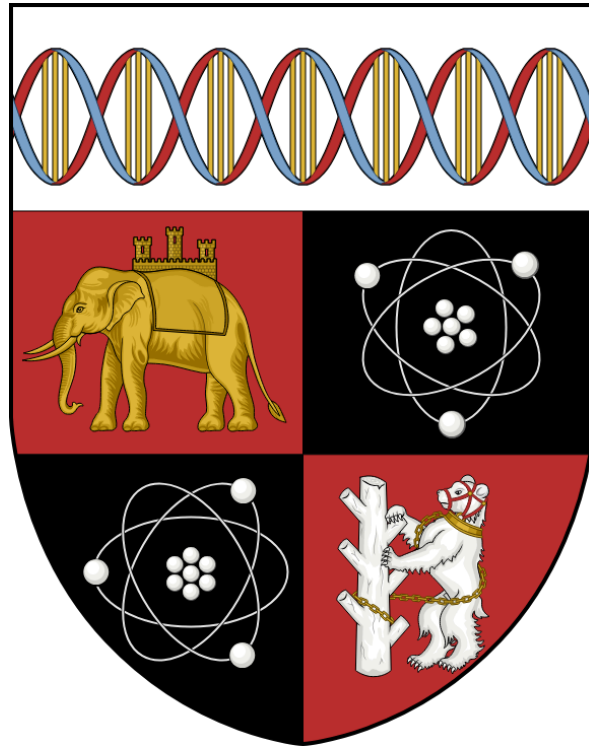


# Explainability in Medical Image Classification



Jacob Davies



# 1 Introduction

## 1.1 Problem statement

Every year, there are over 300,000 new brain cancer cases; brain tumours are among the most expensive cancers per patient due to the combination of complex surgery, ICU care, radiotherapy, chemotherapy, recurrence, and palliative care (Ilić and Ilić, 2023; Johannes Pöhlmann et al., 2024). Brain tumours benefit hugely from early detection and intervention, which makes treatment less intensive and lower risk (Liam MacLuan-Hodgson, 2025). Diagnosing tumours relies on MRI imagery; interpreting this imagery requires specialist neuroradiologists, resulting in slower turnaround in settings lacking sufficient neuroradiology expertise (European Society of Radiology, 2010; The Royal College of Radiologists, 2023). Furthermore, access to trained personnel is uneven globally, with low and middle-income countries facing significant shortages in radiology workforce capacity (Mollura et al., 2020; World Health Organization, 2020). ML and DL vision models have been shown to accurately screen, interpret, and diagnose MRIs; however, regulatory requirements for SoMD (Software as a Medical Device) from MHRA and NICE, explicitly set out the requirement for transparency, explainability, and human oversight, similar requirements are set out by the EU AI act and FDA guidance (European Union, 2024; FDA, 2024; GOV.UK, 2021; GOV.UK, 2025).

A problem with DL is that, despite its benefits, its black-box nature makes regulatory approval difficult to acquire and diminishes its usefulness as a diagnostic aid; as clinicians still need to understand why a diagnosis is being made to make the diagnosis themselves. This project aims to focus on this by evaluating the applicability of a range of explainability tools and evaluating the difference in the quality of explanations for different model architectures.

## 1.2 Dataset

We will use a 4-class brain MRI dataset (Masoud Nickparvar, 2026), which contains 7,200 human brain MRI images 1400 for each class in the training data and 400 for each class in the test. The 4 classes are glioma, meningioma, pituitary, and no-tumour.

## 1.3 Methodology

This is a brief overview of our methodology.

We will evaluate a CNN, a vision transformer, and two ML baselines: HOG+SVM and PCA+RF. We will compare all four models on the four-class brain MRI dataset. The dataset is split into training, validation, and test sets, with the test set held until final evaluation.

Classification performance is assessed using accuracy, macro F1, macro recall, and tumour recall, with confusion matrices used to investigate misclassifications in more detail. To evaluate explainability, we will first apply occlusion sensitivity across all four models as a fair comparison. We will additionally apply a model-native explanation to each architecture to gain deeper insight into its internal reasoning.

Based on the model performance and our interpretation of explainability, we will evaluate the model's applicability as a potential solution.

Finally, we will be looking at how this research compares to historic, recent and future research.

## 2 Models and design decisions

Let's briefly look at the ML and DL models before critically evaluating and explaining the key differences between traditional machine learning and deep learning approaches.

### HOG + SVM

A histogram of Oriented Gradients feature extraction with a Support Vector Machine classifier. This system works by dividing the image into local 8x8 pixel cells, computing gradients within each at different orientations, and aggregating them into histograms across discrete orientation bins; cells are grouped into blocks and normalised (L2-Hys) to reduce sensitivity to illumination variation. (Cortes and Vapnik, 1995; Dalal and Triggs, 2005). A descriptor is used to capture edges and shape boundaries; this approach requires the engineer to decide in advance which gradients encode the relevant signal. As we don't have domain-specific knowledge, we use standard settings from the original literature (Dalal and Triggs, 2005).

### PCA + RF

Principal Component Analysis is used to reduce the 4,096-dimensional input (64x64 pixel brightness values) to 100 principal components. These are then classified with a random forest (Breiman, 2001; Pearson, 1901). The features are standardised before PCA, ensuring components are not dominated by high-intensity pixels. Unlike HOG, this approach utilises no domain knowledge whatsoever, as PCA maximises variance, which is not necessarily equivalent to relevant structures. This serves as a lower-bound baseline, testing whether structure exists in pixel space.

### CNN

Convolutional Neural Networks learn hierarchical features through layers of filters, from low-level edges to higher-level spatial structures. We can adjust the number of filters and layers to enable the model to capture more features and more complex ones. Adjusting kernel size and shape allows larger features to be modelled, and features to be modelled more granularly. Granularity can also be adjusted using the step size of the features (O'Shea and Nash, 2015).

For our system, we are fine-tuning a pre-trained model (specifically, the PyTorch built-in Resnet-18) (Pytorch.org, 2024). Fine-tuning is when you take a model that has already been trained and do a small amount of training on task-specific data, reducing training cost and improving model reliability by

leveraging pre-trained, meaningful features. In theory, the model utilises more general features that it learned from the pre-training, reducing overfitting, reducing performance drop in testing and deployment. ResNet-18 was chosen over deeper variants (ResNet-50, ResNet-101) because it is lightweight, its skip connections reduce vanishing gradients without needing very deep tuning, and it has well-documented transfer learning performance on medical imaging benchmarks.

Choosing to fine-tune decides many of your hyperparameters for you, as the same shape has to be used for the model as is provided; however, here are some key hyperparameters that we still control:

| Hyperparameter | Value             | Rationale                                                                                                       |
|----------------|-------------------|-----------------------------------------------------------------------------------------------------------------|
| Learning rate  | 1e-4              | Deliberately low for full fine-tuning, avoids catastrophic forgetting.                                          |
| Weight decay   | 1e-4              | Light L2 regularisation; appropriate for a relatively shallow model                                             |
| Epochs         | 17                | Chosen empirically; validation loss plateaus around epoch 15                                                    |
| Batch size     | 64                | Fits in memory, larger batches stabilise gradient estimates                                                     |
| Optimizer      | Adam              | Adaptive per-parameter learning rates suit fine-tuning, where different layers need different update magnitudes |
| Scheduler      | ReduceLROnPlateau | Halves the learning rate when validation loss stagnates, preventing oscillation in later epochs                 |
| Image size     | 128×128           | Trade-off between resolution and training speed; ResNet-18 is resolution-agnostic                               |

When fine-tuning, it is possible to run into ‘catastrophic forgetting’ where the model overfits on task-specific data. This would become apparent with a difference in the model performance on training and validation data. We reduce this by starting at a low training rate, but if this were to still become a problem, we could freeze the weights of the earlier layers, forcing the model to use the pre-learned features and only train the last layers.

## ViT

Vision Transformers divide an image into fixed patches (vectors where each entry is a brightness value for a colour channel in a pixel in the patch) and process the patches using a series of self-attention mechanisms and feed-forward networks with normalisation steps within (Dosovitskiy et al., 2020). The size and spectrality of patches, the number of transformer blocks (attention - feed-forward-network layers) and the number of heads can be adjusted to learn more complex relationships or model more relationships. ViTs are very large models and generally expensive to train as size scales quadratically with patch count; as such, we also utilise fine-tuning a pre-trained model (specifically PyTorch vit\_tiny\_patch16\_224) to reduce computational costs (Huggingface.co, 2020).

ViT-tiny was chosen specifically because it is the smallest standard ViT available in timm (PyTorch Image Models) that still uses standard 16×16 patches on 224×224 images, making it computationally

feasible for this coursework. The model is comprised of 12 transformer blocks with 3 attention heads each. The hyperparameters that we can tune and there rational are:

| Hyperparameter | Value             | Rationale                                                                                    |
|----------------|-------------------|----------------------------------------------------------------------------------------------|
| Learning rate  | 1e-4              | Deliberately low for full fine-tuning.                                                       |
| Weight decay   | 1e-2              | Higher than the CNN; AdamW with higher weight decay is standard for transformer fine-tuning. |
| Epochs         | 6                 | Chosen empirically; validation loss plateaus around epoch 4/5                                |
| Batch size     | 32                | Smaller than CNN due to the 224×224 input size and higher memory per forward pass            |
| Optimizer      | AdamW             | Decoupled weight decay is standard for transformers                                          |
| Scheduler      | CosineAnnealingLR | Smooth decay to near-zero by the final epoch.                                                |

## 3 Results

How do these different models compare? To fairly evaluate the models, we are going to evaluate their classification performance on test data using 4 different metrics: accuracy, macro recall (mean class recall), macro F1 (mean class F1), and tumour recall (recall on ‘not no tumour’, we added this to filter out poor performance from similar-looking tumours). Additionally, we will analyse their behaviour by looking at the confusion matrices. It is well documented that DL models struggle to distinguish glioma and meningioma due to visual similarities; as such, it's worth further evaluating the models on their performance for this metric (Gorenshtein, Liba and Goren, 2025).

### Performance metrics

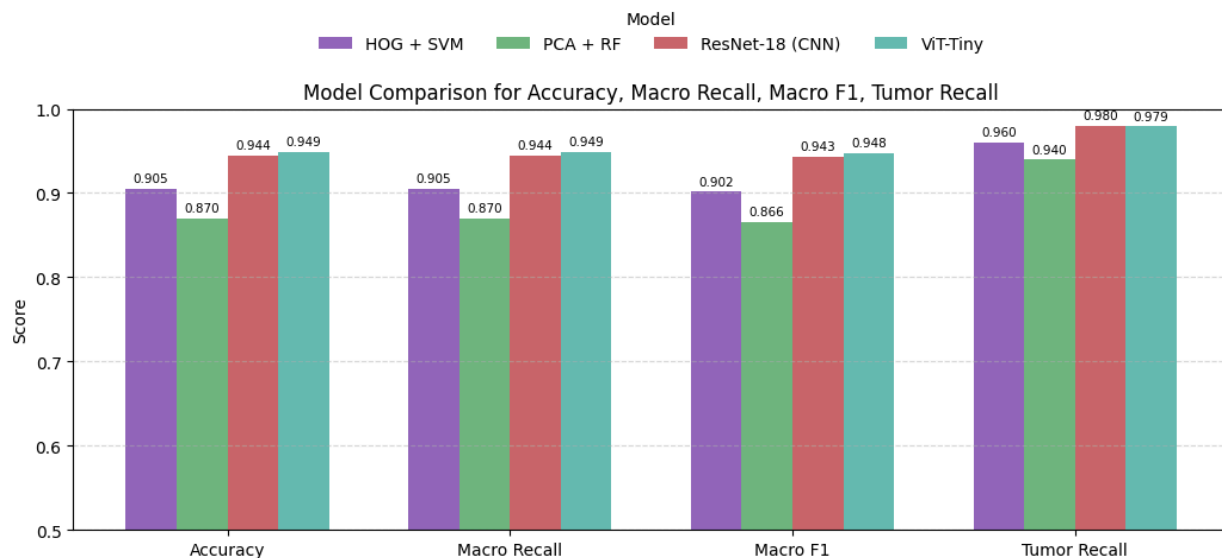


*Figure 3.1 - All 4 performance metrics for all 4 models*

Figure 3.1 shows that there is a clear pattern in performance, with PCA+RF performing the worst, HOG+SVM performing better, and the two deep learning models performing better again.

The performance gap between traditional ML and DL is significant. Deep learning's ability to learn task-specific representations from data enables it to capture more complex patterns in the data, which is especially useful in situations like this, where the vast majority of tumours are correctly being classified, and discrepancies in performance are being caused by a handful of the most difficult cases.

Within DL, the vision transformer system outperforms the CNN on most metrics. This result is consistent with findings in the literature, transformer-based architectures can model global spatial relationships through self-attention, enabling wider, more context-aware feature representations. However, despite the ViT's performance, on tumour recall, it is beaten by a small margin (0.001), which is noteworthy, as on other metrics, it beats the CNN by 0.005. Given that context, a 0.006 performance metric drop becomes more notable. While the performance is still good, it's interesting that it is not entirely uniform; allow us to look at the confusion matrices to try to understand why that's happening.

## Confusion matrixes

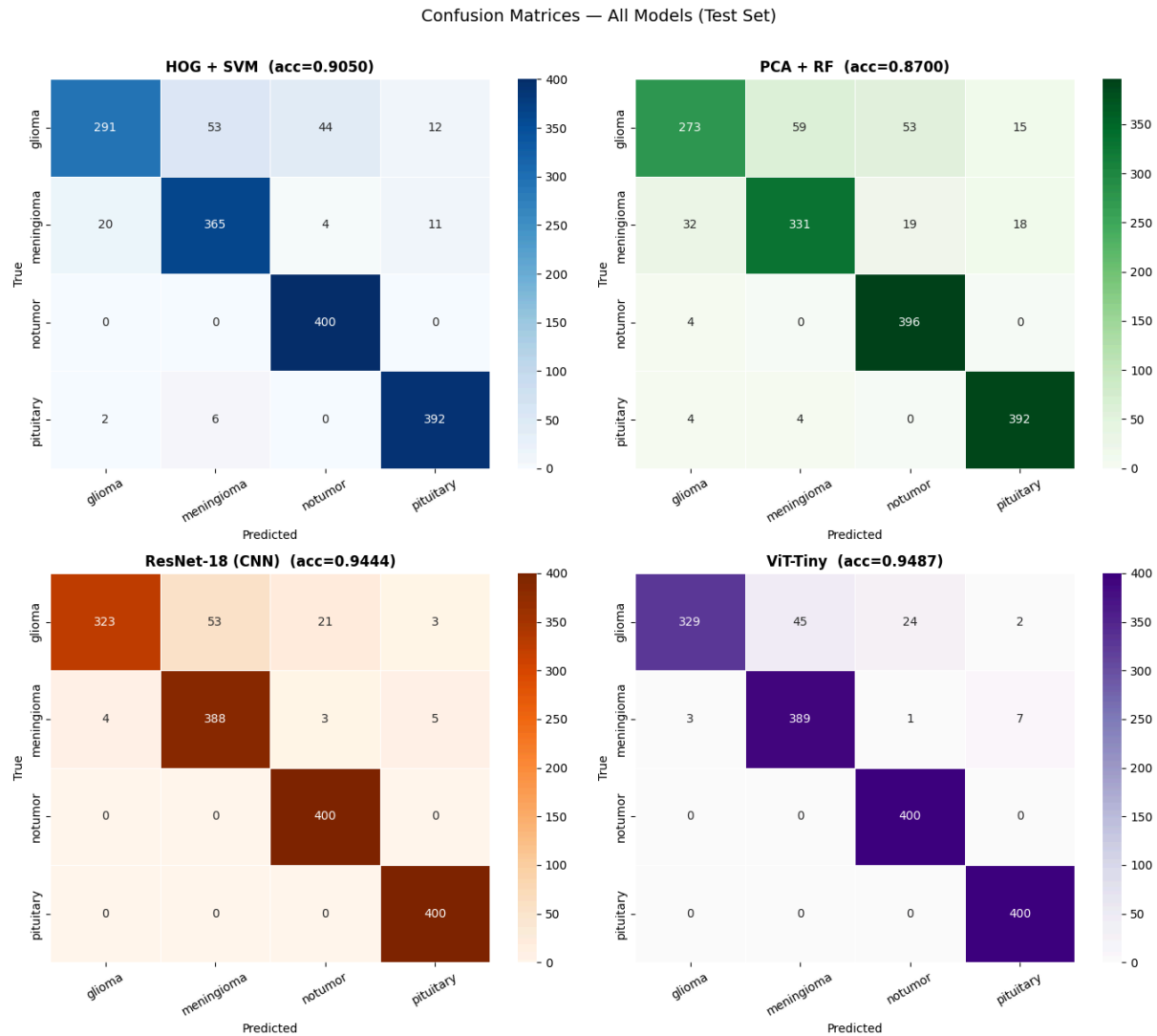


Figure 3.2 4 class confusion matrices, for all 4 models

Looking into this performance drop further, this is driven by missing tumour classifications on glioma tumours, which make up the vast majority of missed tumours in the DL models (97% of missed classifications from ViT, and 87% from CNN). We reason that this is likely due to the fact that gliomas, especially a certain type of glioma called an LGG (Low-grade glioma), infiltrate surrounding tissue especially well, meaning that they have low contrast on MRIs and weak boundaries (Larsen et al., 2017). This may mean that CNNs work better as they can develop filters that are better at picking up weak gradients. While we don't think the numbers are statistically significant, the fact that certain models perform better at different tasks could be important later on, as a model that can do what humans struggle with is more useful as a diagnostic aid.

Something that we touched on a bit earlier was the model's ability to differentiate glioma and meningioma, being documented as a difficult task; we can see this reflected in the confusion matrices,

with these making up the most common misclassifications (Gorenshtein, Liba and Goren, 2025). To quantify how this changes with different models, we can use the glioma-meningioma Confusion Rate as a metric.

$$\frac{(\text{Meningiomas as Glioma Count}) + (\text{Gliomas as Meningioma Count})}{\text{True Glioma} + \text{True Meningiomas}}$$

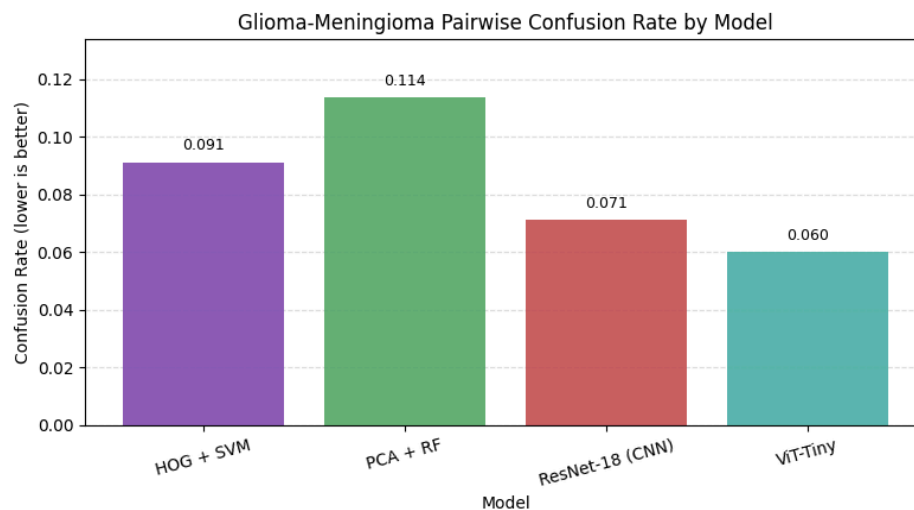


Figure 3.3 - glioma-meningioma Confusion Rate for all 4 models

We can see from this that the quality (lower is better) in model performance follows the same pattern as the general performance metric, likely for the same reasons as mentioned above.

Additionally, the misidentification is asymmetric, with models much more likely to misidentify glioma as meningioma compared to meningioma as glioma. This is consistent with our prior findings of difficulty diagnosing glioma, and our reasoning about its low contrast and weak boundaries (Larsen et al., 2017).

## Explainability

While the high accuracy and recall of our models are good, it isn't the entire story; for a model to be useful, it also has to be explainable. We can use explainability tools to help achieve this. We want to use these tools to make the model's predictions understandable, but it's also useful to ensure that the model is learning real features to increase confidence that the model will work well with real data.

We will be using two explainability tools for each model. We will use occlusion sensitivity for all models, as it allows us to make a fair comparison between models and allows us to assess the model's ability to learn real features. We will also look at a native explanation for each model, which allows us to get a better insight into how the model is working with more accurate and interpretable results.

We will begin with occlusion sensitivity, which works by sliding a grey patch across the input and measuring the drop in correct classification probability. This system works for all models and works in the same way, making it a fair comparison. Below is a grid showing this for all 4 models.

Occlusion-Sensitivity Heatmaps — Why each model classified this image

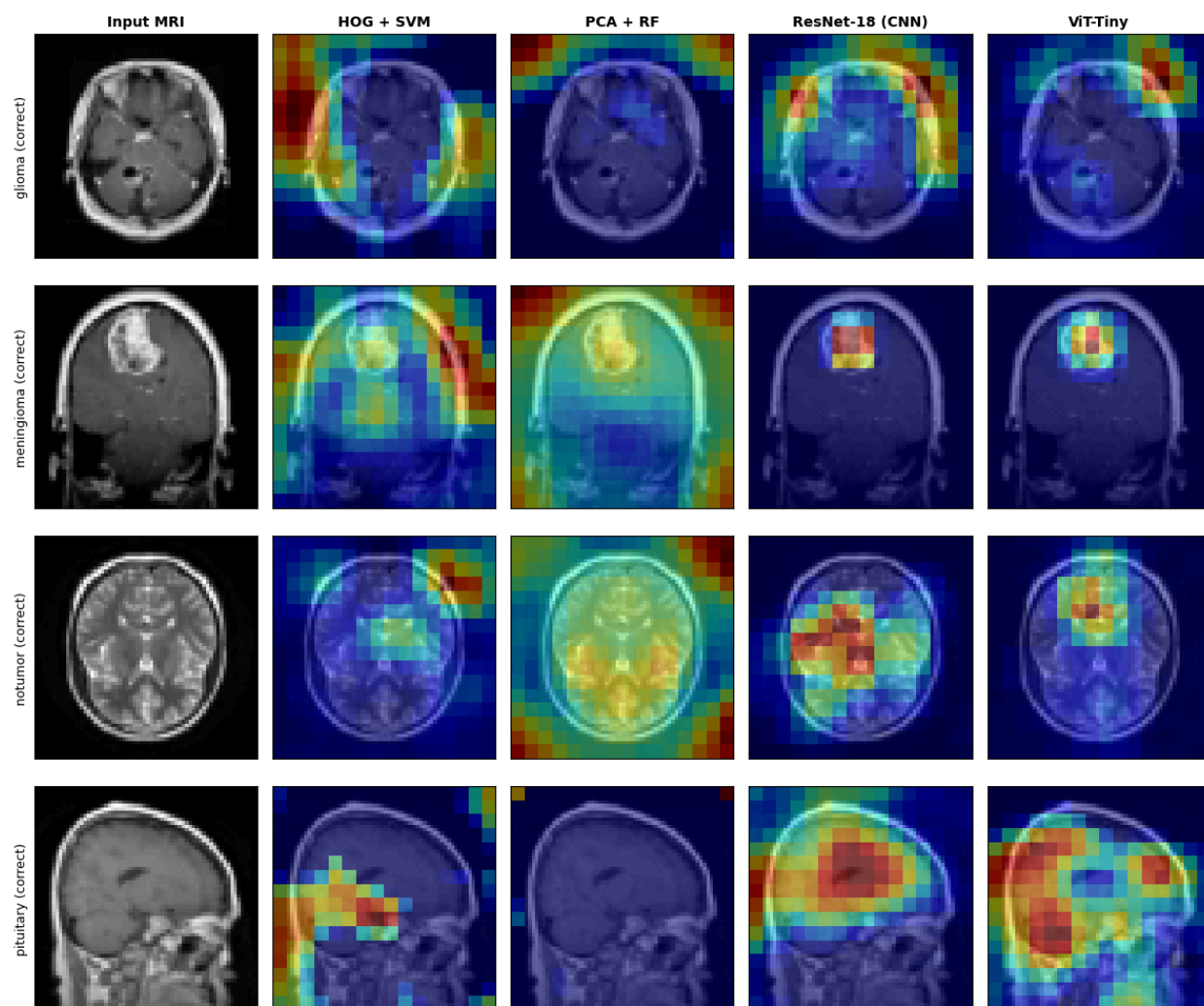


Figure 3.4 occlusion sensitivity heatmaps for 4 MRIs for all 4 models

These results clearly show why the ML models perform worse than the DL models; the ML models primarily rely on features like the brain edges and skull, while the DL models focus on smaller aspects of the brain, around tumours; you can see this clearly in the meningioma MRI, with PCA+RF focusing on the entire image, HOG+SVM looking at a large patch - especially the skull, CNN focusing only on the tumor, and ViT focusing on the middle of the tumor.

These results suggest that DL models are more interpretable, helpful to clinicians, and make predictions based on the real tumour data in the MRI, making them more resilient in deployment with new data.

For the native explanations, we will use a different explainability metric for all 4 models.

- HOG+SVM: 8×8 cell gradient magnitudes (this is the output of HOG, not the SVM attribution). We will further look at the HOG gradient-orientation visualisation.
- PCA+RF: RF feature importance back-projected through the PCA components (this shows global importance, not per-image).
- ResNet-18: Grad-CAM on layer 4 - because it is the deepest convolutional stage, before spatial information is collapsed by global average pooling
- ViT-Tiny: attention rollout across all 12 blocks.

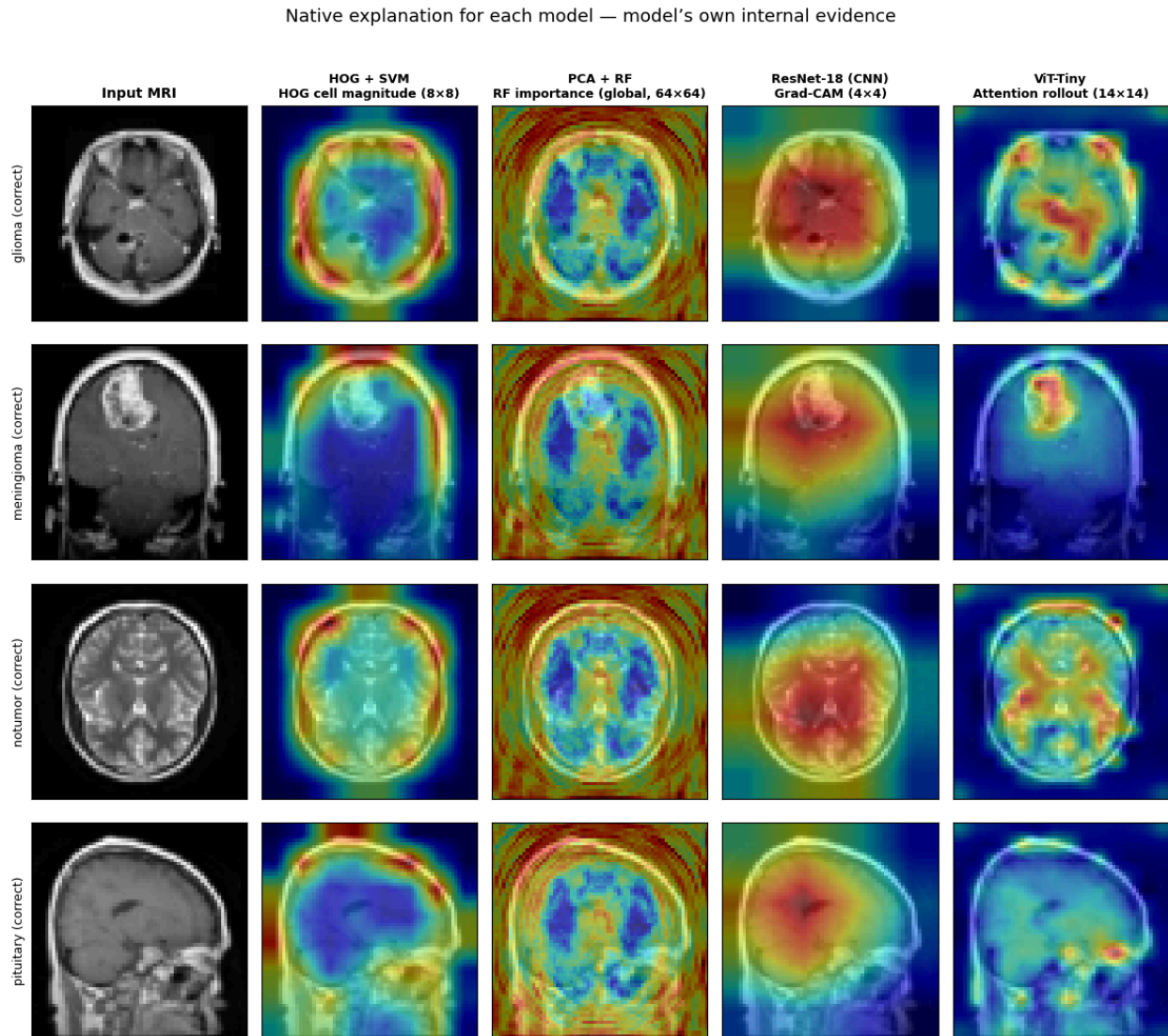


Figure 3.5 - native explanations as heatmaps (RF importance is global; these outputs are from different metrics and are not directly comparable)

The results indicate that we get similar results to the occlusion mapping. Focusing on the meningioma again, ignoring PCA for now, HOG+SVM focuses almost entirely on the skull, whereas the DL models focus on the tumour. It is also visible that the CNN focuses on a broad area, while the ViT focuses very precisely on the tumour.

However, for HOG, there is additional information that we can attain; we can not only look at the significance of cells, but also what gradients are causing said significance. This is a standard metric used for HOG and allows us to utilise the extra explainability that comes with an ML model.

Histogram of Oriented Gradients — what the HOG+SVM model sees

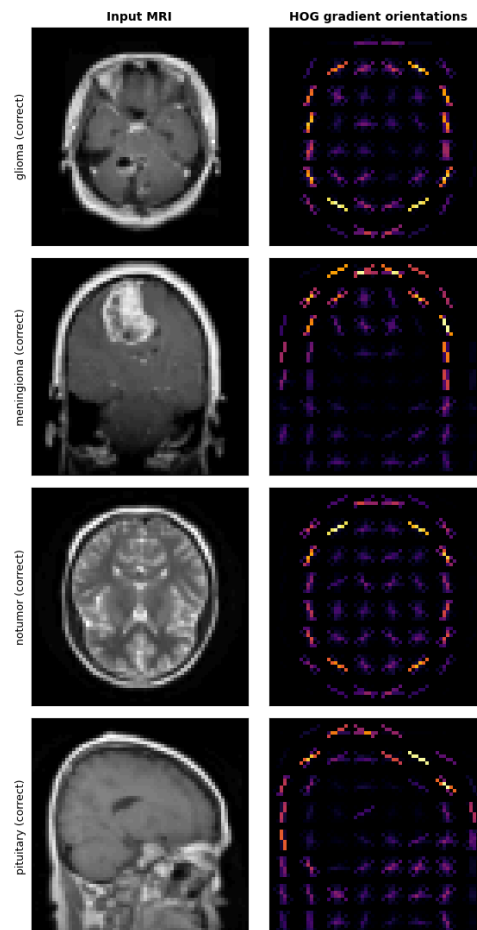


*Figure 3.6 - HOG direction and magnitude of gradients*

This reveals one of the issues with this explainability technique: it is incredibly difficult to convert this additional explanation into useful information.

In general, the above two explainability visualisations present the same conclusion: DL is learning the important features required to identify tumours, increasing confidence in their ability to work on real-world data, and making them useful in assisting clinicians.

## 4 Discussion and future impact

### 4.1 Regulatory applicability

This project focuses on the requirements set out by the EU AI Act, FDA, and MHRA/NICE requirements for SoMD (Software as a Medical Device) by making the models we have trained transparent and auditable.

The dataset used here has been pre-processed (all images contain a cross-section that is diagnosable) and is not drawn directly from a clinical pipeline, so real-world performance would need to be validated with real-world data before any deployment could be considered. CE marking clearance, for example, would additionally require continuous monitoring of performance and audit trials (European Union, 2017).

We propose that the use of a system like this is not as an autonomous diagnosis tool but as a CAD (clinician-assisted diagnosis) tool. Despite the lower regulatory bar, it also reduces the cost of misdiagnosis, as a second set of eyes assesses all scans. The explainability tools explored are excellent for this, as they help a clinician quickly see what aspects of an MRI are important in making decisions, reducing latency in diagnosis and subsequently increasing clinician throughput.

### 4.2 Limitations of explainability

The explainability visualisations have not been validated against expert annotation or ground-truth segmentation masks, so model explanation remains unconfirmed; the models may not be using entirely different features to make diagnoses. If a plausible-looking heatmap does not correspond to actual tumour tissue, this could give a clinician false confidence in an incorrect prediction.

Unfortunately, the resources required to alleviate these concerns far outweigh the scope of this project. A next step would be to test a dataset with a segmentation mask, and then validate the results with a medical professional.

### 4.3 Future trends

The recent shift from the use of CNNs to ViT in medical imaging shows that the advantages offered make the tradeoffs worth it (Aburass et al., 2025). However, there is still much research that needs to be done; it is undecided whether the benefits of general-purpose models make them more effective than domain-specific architectures. Furthermore, emerging approaches such as 3D transformer-based models present new research opportunities by looking at the volumetric brain structure as opposed to a slice.

Future research should focus on determining these unknowns; furthermore, regulations should be continuously updated to allow for the best quality tools to be able to help patients while ensuring patient safety.

## 5 Conclusion

In this project, we have thoroughly and critically evaluated and explained the key differences between traditional machine learning and deep learning approaches by looking in depth at the differences in architecture, implementation, performance, and explainability. We went on to use this lens to evaluate the implications of AI and DL solutions. Finally, we evaluated emerging trends and research development in AI throughout, and used this to ground our research and future developments.

## Bibliography

Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M. and Al-Kadi, O. (2025). Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases. *Journal of Imaging Informatics in Medicine*. doi:<https://doi.org/10.1007/s10278-025-01481-y>.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp.273–297. doi:<https://doi.org/10.1007/BF00994018>.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, 1, pp.886–893. doi:<https://doi.org/10.1109/cvpr.2005.177>.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*. [online] Available at: <https://arxiv.org/abs/2010.11929>.

European Society of Radiology (2010). The Future Role of Radiology in Healthcare. *Insights into Imaging*, [online] 1(1), pp.2–11. doi:<https://doi.org/10.1007/s13244-009-0007-x>.

European Union (2017). *EUR-Lex - 32017R0745 - EN - EUR-Lex*. [online] [eur-lex.europa.eu](http://eur-lex.europa.eu). Available at: <https://eur-lex.europa.eu/eli/reg/2017/745/oj> [Accessed 22 Apr. 2026].

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. [online] [Europa.eu](http://eur-lex.europa.eu). Available at: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.

FDA (2024). Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. *FDA*. [online] Available at: <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles> [Accessed 20 Apr. 2026].

Gorenshtein, A., Liba, T. and Goren, A. (2025). Lightweight Transfer Learning Models for Multi-Class Brain Tumor Classification: Glioma, Meningioma, Pituitary Tumors, and No Tumor MRI Screening. *Journal of Imaging Informatics in Medicine*. doi:<https://doi.org/10.1007/s10278-025-01686-1>.

GOV.UK. (2021). *Software and AI as a Medical Device Change Programme*. [online] Available at: <https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme>.

GOV.UK. (2025). *Software and Artificial Intelligence (AI) as a Medical Device*. [online] Available at: <https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device>.

Huggingface.co. (2020). *timm/vit\_base\_patch16\_224.orig\_in21k · Hugging Face*. [online] Available at: [https://huggingface.co/timm/vit\\_base\\_patch16\\_224.orig\\_in21k](https://huggingface.co/timm/vit_base_patch16_224.orig_in21k) [Accessed 16 Apr. 2026].

Ilić, I. and Ilić, M. (2023). International Patterns and Trends in the Brain Cancer Incidence and mortality: an Observational Study Based on the Global Burden of Disease. *Heliyon*, [online] 9(7), pp.e18222–e18222. doi:<https://doi.org/10.1016/j.heliyon.2023.e18222>.

Johannes Pöhlmann, Weller, M., Marcellusi, A., Grabe-Heyne, K., Krott-Coi, L., Rabar, S. and Pollock, R.F. (2024). High costs, Low Quality of life, Reduced survival, and Room for Improving treatment: an Analysis of Burden and Unmet Needs in Glioma. *Frontiers in Oncology*, 14. doi:<https://doi.org/10.3389/fonc.2024.1368606>.

Larsen, J., Wharton, S.B., McKeivitt, F., Romanowski, C., Bridgewater, C., Zaki, H. and Hoggard, N. (2017). ‘Low grade glioma’: an update for radiologists. *The British Journal of Radiology*, 90(1070), p.20160600. doi:<https://doi.org/10.1259/bjr.20160600>.

Liam MacLuea-Hodgson (2025). *Breaking down new evidence to support a faster, better diagnosis*. [online] The Brain Tumour Charity. Available at: <https://www.thebraintumourcharity.org/news/policy-news/breaking-down-new-evidence-to-support-a-faster-better-diagnosis> [Accessed 21 Apr. 2026].

Masoud Nickparvar (2026). *Brain Tumor MRI Dataset*. [online] Available at: <https://doi.org/10.34740/kaggle/dsv/14832123> [Accessed 13 Apr. 2026].

Mollura, D.J., Culp, M.P., Pollack, E., Battino, G., Scheel, J.R., Mango, V.L., Elahi, A., Schweitzer, A. and Dako, F. (2020). Artificial Intelligence in Low- and Middle-Income Countries: Innovating Global Health Radiology. *Radiology*, 297(3), pp.513–520. doi:<https://doi.org/10.1148/radiol.2020201434>.

O’Shea, K. and Nash, R. (2015). An Introduction to Convolutional Neural Networks. *arXiv:1511.08458 [cs]*, [online] 2. Available at: <https://arxiv.org/abs/1511.08458>.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), pp.559–572. doi:<https://doi.org/10.1080/14786440109462720>.

Pytorch.org. (2024). *resnet18 — Torchvision main documentation*. [online] Available at: <https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html> [Accessed 18 Apr. 2026].

The Royal College of Radiologists (2023). *Clinical Radiology Workforce Census 2023*. [online] Available at: <https://www.rcr.ac.uk/media/5befglss/rcr-census-clinical-radiology-workforce-census-2023.pdf>.

World Health Organization (2020). *Global strategy on human resources for health: Workforce 2030*. [online] WHO. Available at: <https://www.who.int/publications/i/item/9789241511131>.

## Appendix

### Appendix A - AI disclosure

AI has been used to assist with specific tasks such as bug fixing, error explanation, research, validating academic integrity, checking spelling and grammar (see cover sheet), and evaluating my work. Any model outputs have been critically evaluated and modified.

### Appendix B - Document details

Word count including quotations, tables, figures, footnotes, endnotes, in-text citations, titles, abstracts & summaries: 3038 words (2800 + 8.500%)