Interpretable Image Classification for Brain Tumour Diagnosis

1 Abstract:

Deep neural networks have demonstrated remarkable performance in many domains, often surpassing human capabilities, yet their black-box nature poses significant challenges for applications requiring transparent decision-making processes. This study aims to evaluate and compare the explainability characteristics of different neural network architectures for brain tumor detection, with the objective of identifying the most transparent and clinically applicable approaches. Through a comprehensive evaluation that tests the accuracy, computational cost and interpretability of 9 distinct models across 6 architectural families, namely ResNet-50, ResNet101, EfficientNetB0, EfficientNetB1, LeVit-256, LeVit-384, CoAtNet-0, ViT small and XCiT small. This investigation discovered that CNN-based architectures, specifically ResNets and EfficientNets, exhibit superior explainability characteristics compared to attention-based alternatives in tumour detection applications mainly due to their compatibility with Grad-Cam and GuidedBP techniques. The analysis further revealed an inverse relationship between architectural sophistication and model interpretability, such that the more complex architectures, notably the XCiT, LeViTs and CoAtNets, demonstrated the most significant interpretability challenges. Furthermore, a novel XAI technique coined the multimodal consensus approach was presented in this study, the technique leverages the collective insights of multiple architectural families and models rather than relying on individual models. This proposed consensus-based method consistently outperformed traditional single-model approaches regarding explainability, offering a promising direction for implementing transparent and explainable AI in clinical settings.

2 Introduction and Background Research:

2.1 Tumour detection

Recent advances in artificial intelligence have enabled machine learning models to achieve diagnostic accuracies that surpass human capabilities, prompting healthcare institutions worldwide to explore implementing these technologies in their clinical practices [1], [2]. This revolutionary movement can be attributed to the recent achievements of AI image classification models, which have demonstrated exceptional efficiency in tumour detection from medical scans, achieving accuracy levels exceeding 99% [3], [4]. In a field where more than 371,000 deaths occur yearly due to incorrect medical diagnoses, Artificial Intelligence technologies hold the potential to save lives by the hundreds of thousands [5], [6].

Despite this large potential, the integration of AI in medical diagnostics faces a significant challenge: the need for explainable decision-making. Medical professionals remain justifiably sceptical of this technological innovation, as any life-critical decision requires comprehensive justification and transparency. The need for explainable decisions becomes even more important when human life is at risk, and it is precisely here where many machine learning models still struggle today, as they do not provide justification to accompany their predictions. Indeed, for these models, learning to identify life-threatening tumours from medical scans is a completely different task from that of providing explanations that are human-interpretable

[7], [8], [9], [10]. This limitation creates a critical barrier in clinical settings. What happens when an AI model outputs a diagnosis that contradicts that of a medical practitioner? In such scenarios, the ability to examine the underlying rationale of machine learning models becomes paramount. Through transparent justification of AI predictions, healthcare providers can systematically evaluate the methodological approach that informs any given diagnostic outcome. This transparency enables targeted additional investigations to address specific areas of uncertainty, ultimately transforming subjective interpretations into objective diagnostic conclusions.

2.2 The Black Box Problem:

Most machine learning models, especially deep neural networks, are unable to explain the rationale behind their outputs. This conundrum is commonly referred to as the "black box problem", and it is the biggest criticism of high-performing neural network architectures [11], [12]. A neural network architecture typically contains millions of parameters arranged in multiple consecutive layers, each layer processing information through non-linear transformations. This non-linearity makes it extraordinarily difficult to trace the precise decision-making path that takes input features to outputs. Thus, deep learning models learn abstract features that may not align well with intuitive human reasoning [13], [14]. For instance, in computer vision tasks such as facial recognition, while human visual processing relies on identifiable anatomical features, neural networks develop complex analytical frameworks that operate on multiple levels of abstraction, analysing geometric relationships, and angular measurements that may not have intuitive meaning to us. The sequential transformations that occur during forward propagation create layers of computational complexity that remain opaque to interpretation, even for experts in machine learning. Consequently, examining a neural network's architecture alone cannot illuminate its decision-making methodology.

2.3 Explainable AI:

In response to the black-box problem, the field of explainable AI (XAI) emerged to offer methodologies for interpreting complex neural network systems [15]. XAI encompasses a range of strategies designed to provide transparency and interpretability to complex AI systems, particularly focusing on deep neural networks. Since our research will focus on tumour diagnosis, we will limit the XAI techniques and model architectures discussed to the ones relevant to image classification.

Gradient-weighted Class Activation Mapping (Grad-CAM) represents a significant advancement in making image classification architectures more transparent. This technique specifically targets convolutional neural network architectures (CNNs) by leveraging the gradients flowing into the final convolutional layer of a model to generate precise localisation maps that highlight regions of the image which were crucial for classification decisions. First, the model computes gradients of the target class score with respect to the feature maps of the final convolutional layer. These gradients undergo global average pooling to generate neuron importance weights that capture the significance of each feature map for the target decision. The weighted combination of forward activation maps then produces a heatmap, which is upsampled and overlaid on the original image to create an interpretable visualisation of the model's attention patterns [16].

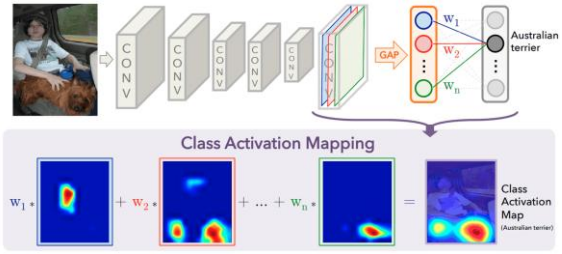


Figure 1: Diagram showing the Grad-CAM methodology workflow [17]

Guided Backpropagation (GuidedBP) offers a complementary approach to CNN interpretation by implementing a modified gradient computation during the backward pass. The technique implements a "guidance" signal during the backward pass that selectively propagates positive gradients while suppressing negative ones. This modification ensures that only features that positively contribute to the classification are highlighted, resulting in clearer and more interpretable visuals [18].

Attention Roll-out, on the other hand, provides a method for visualising how attention propagates through the layers of transformer networks. This technique implements a recursive multiplication procedure that begins at the final layer and works backwards through the network. The process accounts for residual connections by incorporating an identity matrix weighted by relevance scores. This generates comprehensive attention maps that reveal how information flows through the network's multiple layers, providing insights into the model's decision-making process [19].

Beyond Attention, a transformer attention visualisation method as well, introduces a hybrid approach to model interpretability by analysing both attention patterns and feature transformations, enabling a more nuanced understanding of model behaviour. By examining not only where the model focuses its attention but also how it processes and transforms features through its architectural layers, Beyond Attention should provide a more complete picture of the model's decision-making process [20].

Consequently, for CNN-based models like ResNets and EfficientNets, Grad-CAM and GuidedBP provide complementary insights: Grad-CAM offers broader localisation information, while GuidedBP reveals fine-grained feature contributions. For attention-based architectures, the combination of Attention Roll-out and Beyond Attention techniques enables a comprehensive understanding of both the attention mechanisms and feature processing pathways. This research implements these XAI techniques systematically across different architectural families to evaluate not only the model’s interpretability capabilities but also their practical applicability in clinical settings where model transparency is critical.

3 Design and Methodology:

3.1 Data Acquisition

A comprehensive dataset of 7023 MRI brain scan images was obtained from Kaggle. There are four distinct categories of images: no tumour, glioma, meningioma, and pituitary. The dataset was pre-split into training and testing sets.

3.2 Computational Resources

Google Collab was used to perform all relevant computations for this project. The platform was chosen due to its ability to leverage powerful GPUs accessible via its cloud system. To ensure methodological consistency and valid model comparisons, all models were trained using the NVIDIA T4 GPU.

3.3 Framework

PyTorch was used as the foundational deep-learning framework of this project, encompassing the entire machine-learning pipeline from data ingestion and augmentation to model development training, evaluation and interpretation. PyTorch's ecosystem includes pre-trained models, which enhanced our development efficiency. The framework's native support for GPU acceleration also facilitated efficient model training. PyTorch also has active community support and widespread adoption in both academic research and industry applications, which provided additional validation for this technical choice.

3.4 Data Preparation and Augmentation

In the initial preprocessing phase, all images were standardised to a 224x224 format. Following this, a data augmentation pipeline was developed specifically for MRI medical images, incorporating multiple transformation techniques to enhance model generalisation capabilities. This decision was supported by established research demonstrating the significant impact of data augmentation on model performance [20], [21]. The concept of data augmentation is to create meaningful variations in the training images while preserving the relevant features, exposing the models to various ways real-world data might differ from the training set.

The data augmentation techniques used in this project include randomly applying rotations up to 10 degrees and translations of up to 10% of image dimensions to simulate variations in patient positioning during image acquisition. Random zoom transformations of up to 10% were applied to account for scaling variations. Gaussian noise was added with a factor of 0.05 to simulate image artefacts. Additionally, elastic deformation was implemented with controlled parameters (alpha=1000, sigma=30, alpha_affine=30) to create realistic anatomical variations while preserving the image topology. Each transformation had a 50% probability of being applied during training, except for the elastic deformations, which were applied with a lower probability of 20%. This was done to ensure that the models were exposed to both the regular training set images and the augmented ones.

Following this, a custom MRIDataset class that inherits from PyTorch's Dataset class was developed to manage two distinct processing streams: training and testing. The training pipeline implemented real-time augmentation transformations with standardised normalisation (mean=0.485, std=0.229). In contrast, the validation and testing pipeline only performed tensor conversion and normalisation without any augmentation, this was done to maintain consistent evaluation conditions. PyTorch's DataLoader was utilised for efficient batch processing through multiple workers.

3.5 Transfer Learning:

Research has shown that the use of pre-trained models for image classification significantly accelerates training time while simultaneously yielding better accuracy scores [22, 23]. For this reason, all models used in this project were pre-trained, either loaded from PyTorch's or Timm's (torch image models) libraries of models.

3.6 Model Architecture Selection:

A diverse range of image classifier architectures were selected for this project. From the PyTorch library, the ResNet-50 and ResNet-101, alongside the EfficientNet variants B0 and B1, were taken. More sophisticated architectures were taken from the Timm library, those models were the Vision Transformer (ViT_small), the lightweight and efficient LeViT models (LeViT256 and LeVit384), the compound attention network CoAtNet0, and the cross-covariance image transformer (XCiT_small). This selection of models represents a comprehensive spectrum of architectural approaches, from traditional convolutional networks to modern transformer-based architectures.

3.7 Training:

K-fold Cross-validation: Regarding the training process, many advanced techniques were incorporated, such as K-fold Cross-validation for hyperparameter optimisation. This training technique works by randomly dividing the dataset into “k” equal-sized segments; in our case, k=5, the dataset was divided into five equal-sized segments and 20% of the training data was used for validation in each fold. Four of the segments are then combined to form the training set, and the remaining segment serves as the validation set. This process will repeat “k” times, a.k.a five times. The performance metrics are then averaged across all five iterations to provide a more reliable estimate of the generalisation capabilities of the model. For each architecture, various hyperparameter combinations were systematically evaluated. These included different learning rate schedulers (plateau and cosine), initial learning rates ranging from 5e-6 to 1e-3 depending on the architecture, patience values of 5, 10, and 15 epochs, and minimum learning rates of 1e-6 to 1e-8.This process helps identify the most robust set of hyperparameters that performed well across different data distributions. After completing the k-fold cross-validation, a final version of the model is trained using the entire training dataset. This final training produces a model that leverages all available training data.

Early stopping: Another technique incorporated during training was early stopping to prevent overfitting from happening. Overfitting occurs when a machine learning model specialises in the specificities of the training dataset to the extent of losing its generalisation capacities. Practically, early stopping involves monitoring the model's performance on the training and validation sets throughout the epochs of the training process. Model training is forced to stop when the training error keeps getting lower while simultaneously, the validation error keeps growing. If this happens for a set number of consecutive training epochs, the current training iteration gets terminated and moves on to the next.

Learning rate scheduling: An important part of the training process was the incorporation of learning rate scheduling with the Adam optimiser variant of gradient descent. This technique dynamically adjusts the learning rate parameter, which controls how much the model updates its weights in response to observed errors. Initially, the learning rate is intentionally set to a high coefficient to facilitate rapid convergence toward promising regions of the parameter space. During these early stages, the gradient descent algorithm can take larger steps in the parameter space without significant risk of missing important optima points. As training progresses, the learning rate is systematically reduced according to a predetermined schedule. A smaller learning rate allows the gradient descent algorithm to make finer adjustments to the model parameters, reducing the risk of overshooting the global optimum, the theoretical point at which all model parameters are optimally tuned

to minimise the loss function. A static learning rate, on the other hand, would either result in slow initial progress if set too low or potential instability in later stages if set too high.

Others: Moreover, cross-entropy served as the loss function, appropriate for multi-class classification tasks, and performance logs that tracked training duration and validation accuracy across all training folds were tracked and kept. Systematic model saving after each fold was also incorporated.

3.8 Evaluation metrics:

Accuracy: Prediction accuracy was assessed via confusion matrices, accuracy scores and F1 scores. The confusion matrix provided a detailed visualisation of model predictions across the four classes by displaying the number of true positives, true negatives, false positives, and false negatives. The overall accuracy metric quantified the proportion of correct predictions across all cases, while the F1 score provided a balanced performance measure across all classes by combining precision and recall.

Computational Speed and Cost: The computational efficiency assessment tracked training duration and parameter count. Training time was tracked to understand the computational resources required for model development, while the number of parameters was documented to evaluate memory requirements.

3.9 Explainable AI Visualisations:

CNN-based Models: To visualise the decision-making processes of the CNN-based models, notably ResNet50, ResNet101, EfficientNetB0 and EfficientNetB1, Grad-CAM (Gradient-weighted Class Activation Mapping) and GuidedBP (Guided Backpropagation) were used. To accommodate architectural differences, separate visualisation implementations were developed for the ResNet and the EfficientNet architectures. The implementation leveraged the Pytorch-grad-cam library while extending it with custom implementations of GuidedBP utilising colour-coding to indicate activation importance levels, with red highlighting the most significant features and blue indicating moderately important regions. Gaussian filtering was applied to enhance visualisation clarity, alongside contrast enhancement for improved feature visibility.

A key innovation in this project is the multi-model consensus approach that combines insights from multiple models to generate a single visualisation map. This system implements an adaptive thresholding approach that considers model confidence levels and inter-model agreement patterns. It employs a weighted averaging scheme that gives more influence to models with higher confidence, while also adjusting threshold levels based on the degree of agreement between models.

Attention-based Models: To visualise the attention of the transformer-based architectures, Attention Roll-out and Beyond Attention were used as they provide insight into how attention propagates through transformer layers. The Attention Roll-out implementation works by registering hooks that capture attention weights from each transformer block. These weights are then aggregated through matrix multiplication to create a cumulative attention map, effectively showing how attention propagates through the network's layers. Normalisation was applied to ensure the final attention maps were properly scaled and interpretable. The Beyond Attention implementation extends this approach by incorporating additional processing steps. It implements a discard ratio parameter (set to 0.9) to filter out low-attention weights, focusing on the most significant attention patterns. The implementation also adds residual connections through an identity matrix and applies normalisation to

maintain mathematical validity. Both the visualisation techniques process the transformer's attention maps to create heatmap overlays, highlighting regions of the input image most influential to the model's decision-making process.

4 Experimental Results and Discussion:

4.1 Prediction Accuracy:

The most accurate model was the visual transformer achieving 99.8% prediction accuracy on the test set, followed by the ResNet101, Levit384, and ResNet50 models, which all achieved the same accuracy score of 99.5%. The LeViT256 and CoatNet0 models, on the other hand, had the lowest performances, with respective accuracy scores of 96.2% and 96.7%.

The analysis also revealed a linear correlation between accuracy and F1 scores, indicating consistent model performance without bias toward any particular class.

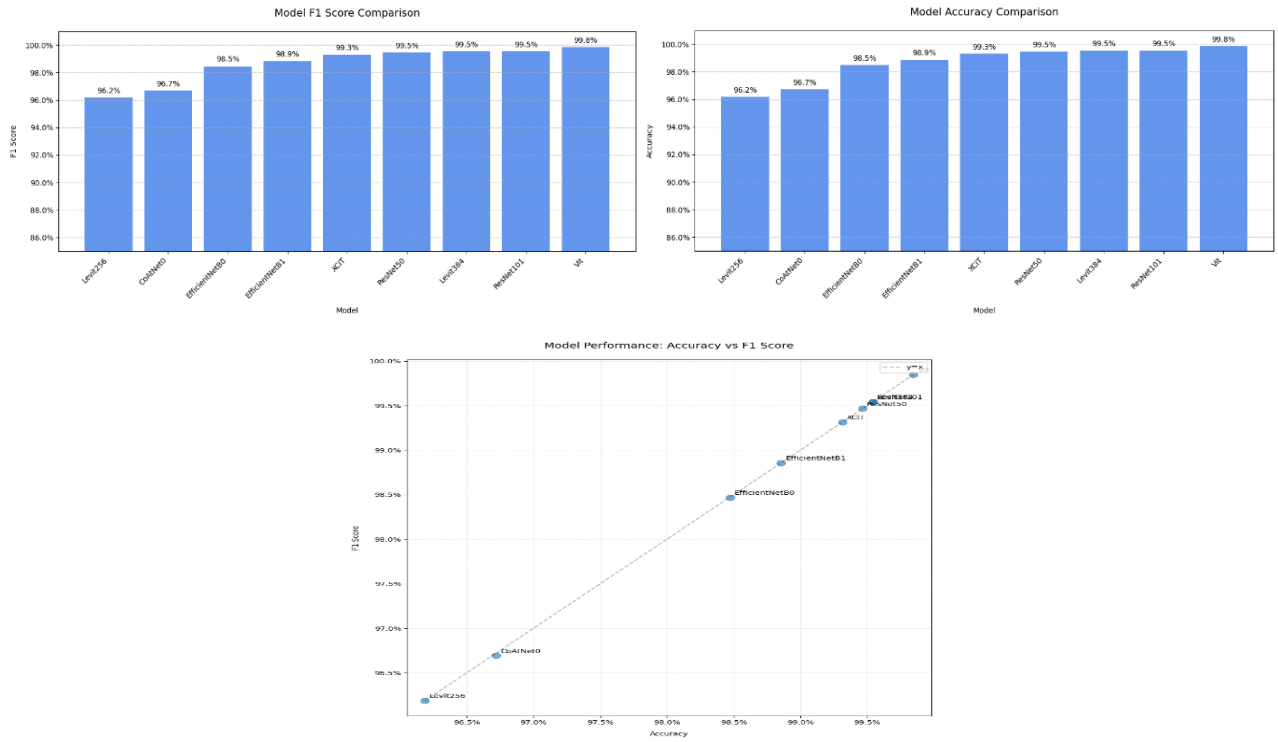


Figure 2: The metric top right figure shows the F1 scores, the top left figure shows the overall accuracy, and the bottom figure demonstrates the linear relationship between overall accuracy and F1.

4.2 Computational efficiency:

The ResNet101 demonstrated the highest computational demands with a training duration of 1.6 hours, followed by the Vision Transformer at 1.4 hours. The EfficientNets B1 and B0, as well as the XCiT, achieved moderate training times of 0.7 hours. Whereas the fastest models to train were the CoAtNet0 and LeViT384 at 0.4 hours only.

Analysis of the relationship between model parameters, training time, and accuracy yielded several notable insights. No correlation was found between either accuracy and training time, nor between the number of parameters and training time. These results seem to point towards the fact that training time is mostly dependent on the model's architectural design itself rather than its number of parameters. However, within the same architectural families, an increase in parameter size generally corresponded with longer training times, with the LeViT384 presenting an interesting exception by training faster than the LeViT256 despite having more parameters. This result could be due to the LeViT's unique specialisation to exploit the parallel processing capabilities of modern hardware reported [24]. Additionally, no significant correlation was found between the number of parameters and accuracy.

The analysis revealed that the EfficientNet models achieved an optimal balance between model size and accuracy. Conversely, the ResNet101 model had the worst accuracy to computational efficiency trade-off of the group. These observations highlight the importance of architectural design in achieving both computational efficiency and high performance in image classification.

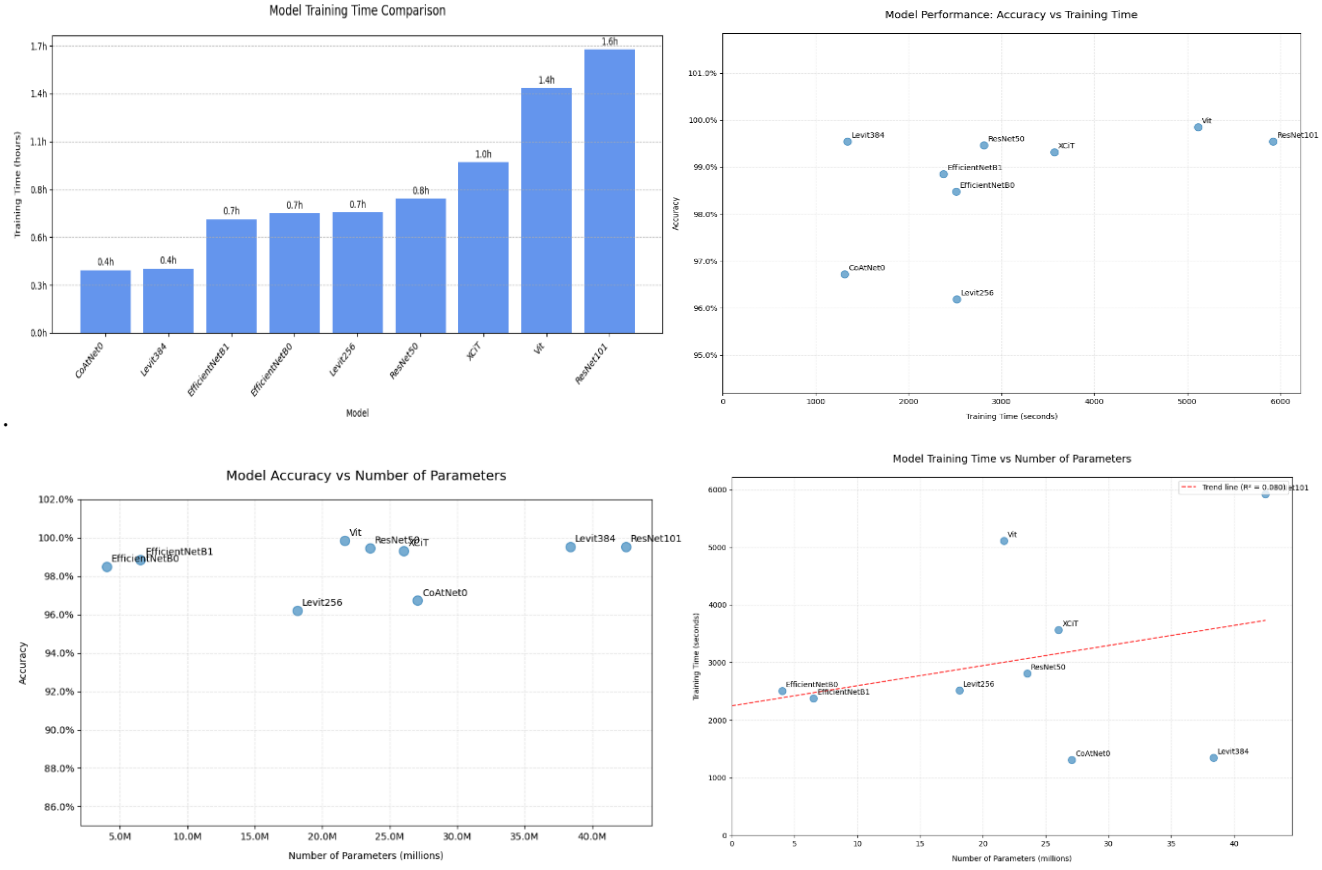


Figure 3: The top left figure shows the training time in seconds of each model, the top right figure compares model performance with training time, the bottom left figure compares training time to the number of parameters, and the bottom right figure shows the relationship between number of parameters and accuracy (the obstructed data point in the top right is the ResNet101).

4.3 Explainable AI Visualisations and The Multi-Modal Consensus Approach:

The implementation of multiple XAI techniques has yielded valuable insights into the distinct ways different model architectures learn to detect and classify tumours. The observed differences in visualisation patterns can be attributed to specific characteristic differences between architectures. EfficientNets produced more localised Grad-CAM visualisations, while showing scattered global attention in GuidedBP visualisations. Conversely, the ResNet50 model generated the most precise GuidedBP interpretations. The EfficientNets' superior performance in Grad-CAM visualisations could be due to their architecture, which combines MBConv blocks and squeeze-and-excitation modules to facilitate effective feature refinement at each network level. The progressive feature compression in their bottleneck structure maintains spatial information more effectively, resulting in more focused activation maps [25]. However, while these architectural elements are efficient for model performance, they make it challenging for GuidedBP to backpropagate gradients through the EfficientNets effectively, resulting in more diffuse attention patterns.

ResNets' outperformance in GuidedBP, on the other hand, can be attributed to the architecture's residual connection mechanism, which facilitates effective gradient flow during backpropagation, leading to more precise feature attribution [24]. These same residual connections, however, can cause gradient diffusion and less localised activation maps in Grad-CAM, where preserving spatial information is crucial for generating meaningful attribution maps that accurately reflect the model's decision-making process.

The multi-modal consensus approach effectively addresses these architectural trade-offs by leveraging the complementary strengths of both EfficientNets and ResNets while mitigating their respective weaknesses. The approach implements adaptive thresholding that considers both model confidence and inter-model agreement, effectively reducing the risk of misinterpretation and increasing the clinical utility of the visual explanations. This systematic integration of visualisation from multiple models enhances reliability by cross-validating findings across different architectural perspectives, compensating for individual architectural limitations while preserving their respective strengths, and thereby providing clinicians with more robust and trustworthy diagnostic insights. This finding suggests that combining multiple architectures may be more effective for the task of explainable tumour classification than the traditional approach of solely relying on a single best-performing architecture.

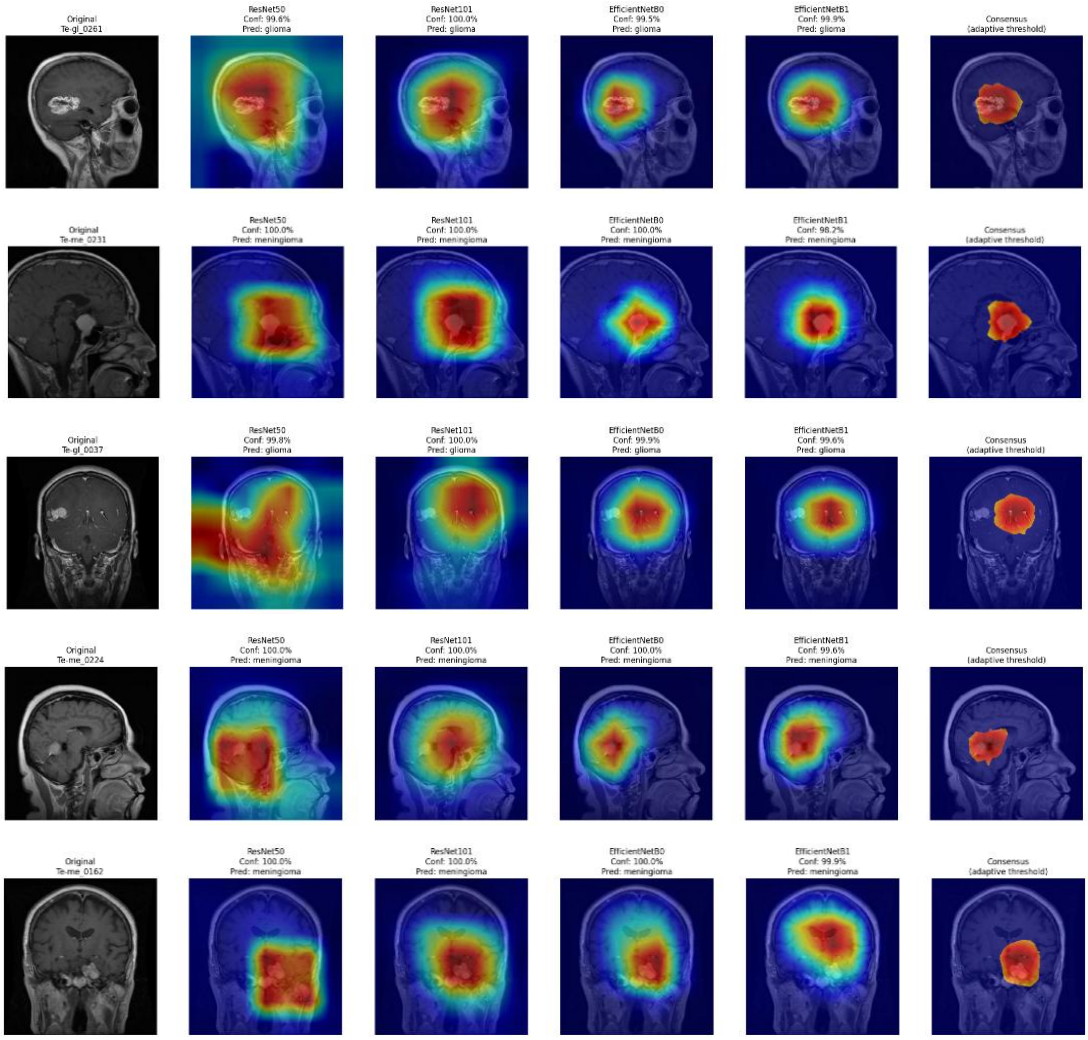
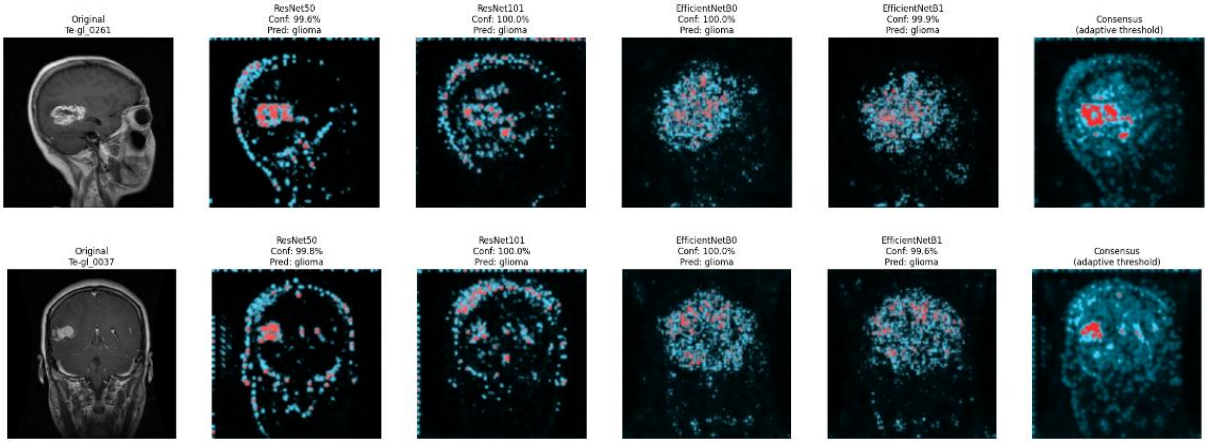


Figure 4: Grad-Cam heatmaps of the ResNet50, ResNet101, EfficientNetB0 and EfficientNetB1 models as well as the combined consensus of all four models, showing areas of overlap between the models



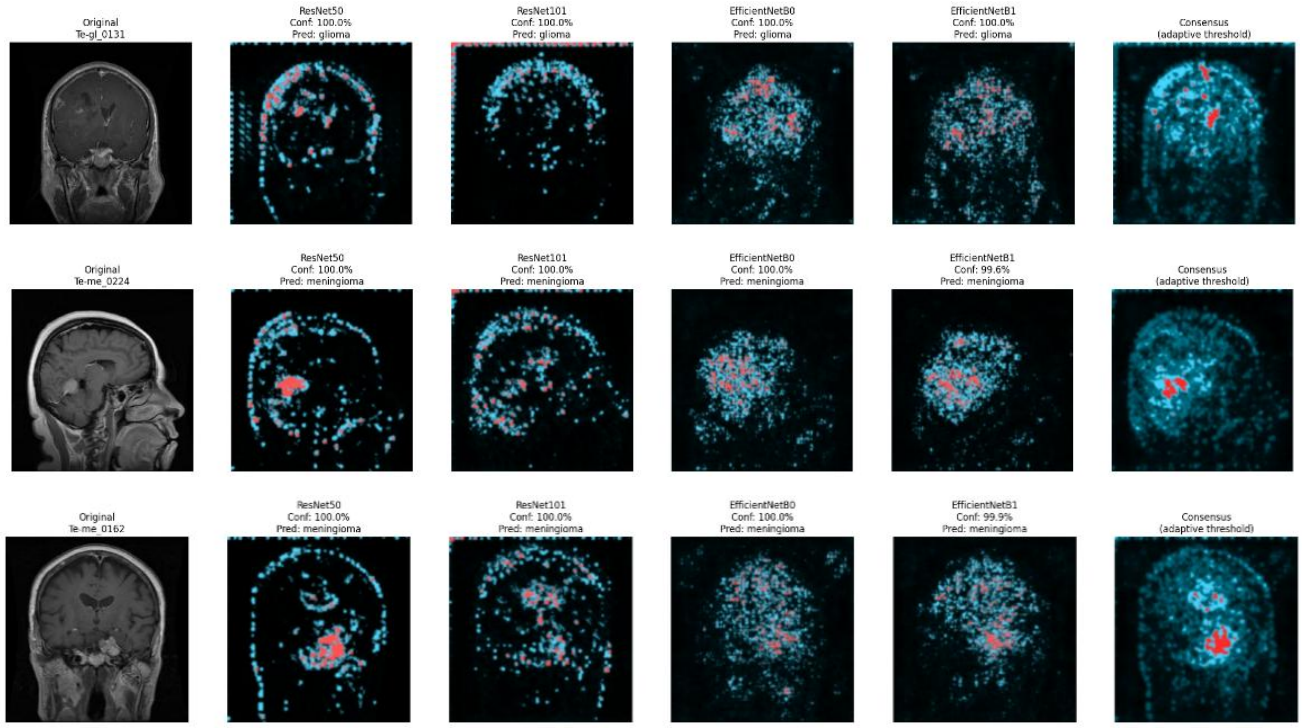


Figure 5: Guided Back-Propagation visualisations of the ResNet50, ResNet101, EfficientNetB0 and EfficientNetB1 models as well as the combined consensus of all four models, showing areas of overlap between the models

The Vision Transformer demonstrated a notable disparity between its classification performance and interpretability. Despite achieving the highest accuracy score of 99.8%, both Attention Roll-Out and Beyond Attention visualisation techniques yielded suboptimal results for interpretation purposes. This discrepancy between the model's high classification accuracy and poor visualisation quality can be attributed to the fundamental architecture of Vision Transformers. While ViT's self-attention mechanism excels at capturing complex relationships between image patches for accurate classification, these same mechanisms make it challenging to generate interpretable visuals. The ViT's patch-based processing and global nature of attention calculations result in diffuse, non-localized feature representations that don't translate well into human-interpretable visualisations. Additionally, the multi-head attention mechanism distributes feature importance across multiple attention heads, making it difficult to aggregate these distributed representations into one coherent image that highlights regions of high interest [26].

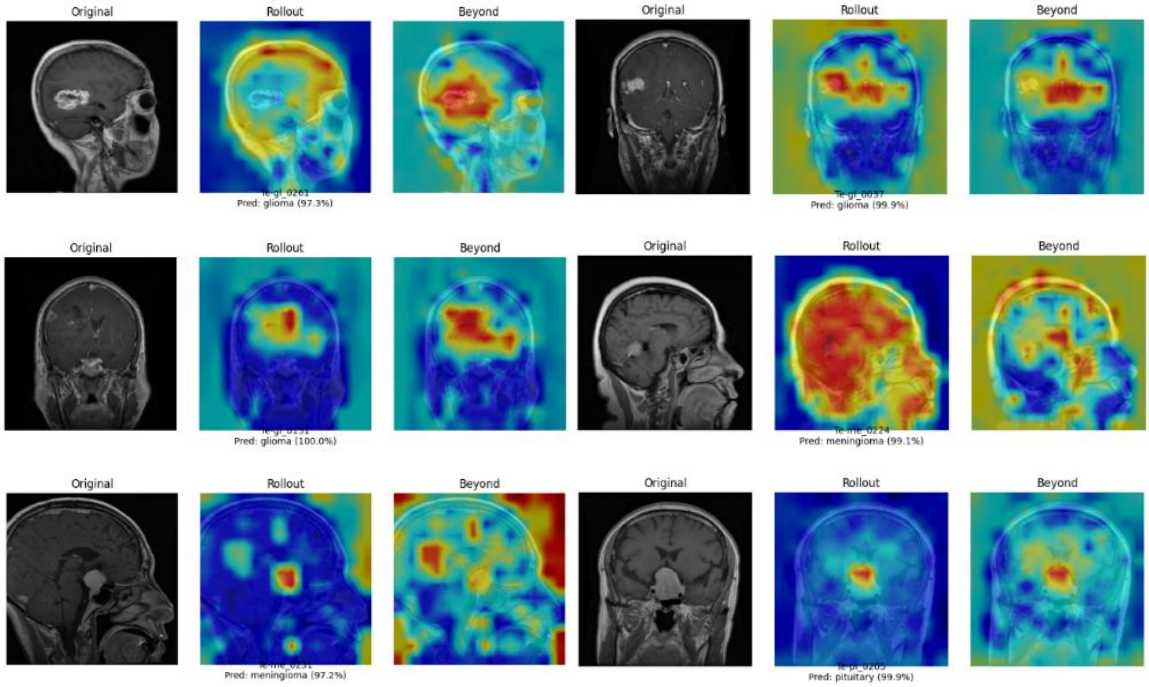


Figure 6: Attention roll-out and Beyond attention visualisations for the ViT Model

4.4 Complex Architectures:

Investigation into the more complex architectures such as the XCiT, and the two hybrid architectures, LeViT and CoAtNett, revealed a significant inverse relationship between architectural sophistication and model interpretability. While these architectures demonstrate remarkable advances in computer vision through improved accuracies and computational efficiencies, their decision-making processes proved particularly challenging to interpret and visualise. The challenge emerged from the incompatibility of these advanced architectures with existing explainable AI frameworks. In the case of the hybrid architectures, the limitation point stemmed directly from the integration of both transformer and convolutional components, rendering the model incompatible with either convolutional-based visualisation methods such as Grad-CAM and GuidedBP or with attention-based ones like Attention Roll-Out and Beyond Attention. The LeViT's two-stage architecture would require substantial modification of traditional attention visualisation methods to handle inter-stage transitions [27]. CoAtNet's alternating convolutional and attention layers created difficulties in capturing information flow between different layer types [28]. And the XCiT's cross-covariance attention mechanism, which operates on feature correlations across channels rather than direct token manipulation, makes standard attention visualisation techniques inadequate [29]. The architectural complexity of these models ultimately diminished the transparency of their decision-making processes, making them more challenging to interpret and ultimately exacerbating the "black box" problem.

5 Conclusion, Future Work and Ethical Issues:

In conclusion, this research demonstrates that the path forward in explainable AI for brain tumour classification lies not in the traditional approach of selecting individual models that achieved the highest performance levels on a testing set, but rather in developing sophisticated frameworks that can leverage the complementary strengths of multiple architectures. Through comprehensive analysis, we demonstrated that no single model or visualisation technique consistently provides optimal interpretability across all scenarios. Instead, it was the integration of multiple models from varying architectural families through the multi-modal consensus approach that was developed in this study, that yielded the most reliable and comprehensive approach for this critical medical task. While the current multi-model consensus approach implementation has demonstrated significant advantages over single-model approaches, there is great potential for further enhancement through the incorporation of more computationally intensive models and the utilisation of a larger pool of models with varying architectures.

The interpretability limitations observed in the LeViT, CoAtNet and XCiT architectures highlight a critical gap in the development methodology of machine learning models that prioritise performance metrics over interpretability. This leads to significant challenges in attempting to understand the decision-making processes of these models, making them unusable in domains such as medical diagnosis, where model transparency is critical for clinical adoption. To address this challenge, future architectural developments should adopt a holistic approach that balances performance improvements with interpretability requirements. This might involve designing built-in interpretation mechanisms to make the models compatible with existing interpretability techniques or developing corresponding interpretability methods for the new architecture. Such approaches would better serve the needs of high-stakes applications where understanding the model's decision-making process is as crucial as its accuracy.

The Interpretability-Aware Vision Transformer (IA-ViT) represents a significant step forward in designing architectures with built-in interpretability. This innovation addresses the fundamental challenge of hybrid architectures by incorporating interpretability considerations directly into the model architecture through a modified attention mechanism that maintains clear attention paths while preserving the high-performance characteristics of traditional Vision Transformers [28]. This architectural innovation paves the way for a new wave of interpretability-aware machine learning architectures, demonstrating that architectural complexity and interpretability do not have to be mutually exclusive.

Lastly, the implementation of AI in clinical settings necessitates careful consideration of ethical implications and regulatory requirements. Healthcare institutions must implement comprehensive validation protocols to assess model predictions across diverse demographics, as architectural variations may exhibit differing performance characteristics across population subgroups. It is also important that there be protocols for cases where model interpretations diverge, ensuring medical practitioners maintain decisive oversight in the diagnostic process. Not to mention that the processing of medical imaging data through machine learning models introduces additional complexity in maintaining patient privacy and data security.

Data set:

https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

GitHub:

https://github.com/samisena/Brain_tumour_classification

Acknowledgements:

I acknowledge the use of the following AI tools in the completion of this work:

Claude 3.5 Sonnet

Publisher: Anthropic

URL: https://claude.ai

This LLM generative AI model was used as a coding assistant to help debug code implementations and suggest potential optimisations for model architectures.

Scipy

Publisher: SciSpace (Typeset)

URL: https://typeset.io

This fine-tuned LLM was utilised as a literature review assistant to help gather and organise relevant research.

References:

1. S., Karthikeyan., S., Lakshmanan. (2024). 1. Image Segmentation for MRI Brain Tumor Detection Using Advance AI algorithm. Available from: 10.1109/otcon60325.2024.10688248
2. Birendra, Kumar, Saraswat., Shashikant, Katiyar., A., Sharma. (2024). 4. Neuro Image Processor For Tumour Detection (NIPTI) using Machine Learning. Journal of advances in science and technology, Available from: 10.29070/62qdeh70
3. A., M., J., Z., Rahman., Muskan, Gupta., S., Aarathi., T., Mahesh., Vinoth, Kumar, Venkatesan., S., Y., Kumaran., Suresh, Guluwadi. (2024). 6. Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering. Available from: 10.1186/s12911-024-02519-x
4. Ashifur, Rahman., Md., Rajaul, Karim., Punam, Chowdhury., Anayet, Hossain., M., Mazharul, Islam. (2023). 7. NeuroXAI++: An Efficient X-AI Intensive Brain Cancer Detection and Localization. Available from: 10.1109/ncim59001.2023.10212818
5. Thornton, C. (2023) 795,000 Americans a year die or are permanently disabled after being misdiagnosed, USA Today. Available at: https://www.usatoday.com/story/news/health/2023/07/18/medical-misdiagnosis-killing-disabling-americans/70423573007/
6. Merelli, A. (2023) Misdiagnoses cost the U.S. 800,000 deaths and serious disabilities every year, study finds, STAT. Available at: https://www.statnews.com/2023/07/21/misdiagnoses-cost-the-u-s-800000-deaths-and-serious-disabilities-annually-study/
7. Emmanuel, G., Pintelas., Meletis, Liaskos., Ioannis, E., Livieris., Sotiris, Kotsiantis., Panagiotis, Pintelas. (2021). 5. A novel explainable image classification framework: case study on skin cancer and plant disease prediction. Neural Computing and Applications, Available from: 10.1007/S00521-021-06141-0

8. Damien, de, Mijolla., Christopher, Frye., Markus, Kunesch., J., Mansir., Ilya, Feige. (2021). 1. Human-interpretable model explainability on high-dimensional data. arXiv: Learning
9. Purushothaman, Natarajan., Athira, Nambiar. (2024). 8. VALE: A Multimodal Visual and Language Explanation Framework for Image Classifiers using eXplainable AI and Language Models. arXiv.org, Available from: 10.48550/arxiv.2408.12808
10. Rami, F., Ibrahim., M., Omair, Shafiq. (2022). 9. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. ACM Computing Surveys, Available from: 10.1145/3563691
11. G., Anushree., Suraj, B, Madagaonkar., C, H, Ravili. (2024). 5. Unveiling the Black Box: A Comprehensive Review of Explainable AI Techniques. Indian Scientific Journal Of Research In Engineering And Management, Available from: 10.55041/ijsrem37405
12. (2022). 6. Explainable AI for Computer Vision. Available from: 10.1007/978-1-4842-8273-1_10]
13. Adityanarayanan, Radhakrishnan., Daniel, Beaglehole., Parthe, Pandit., Misha, Belkin. (2022). 7. Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features
14. Quentin, RV., Ferry., Joshua, Ching., Takashi, Kawai. (2023). 2. Emergence and Function of Abstract Representations in Self-Supervised Transformers. arXiv.org, Available from: 10.48550/arxiv.2312.05361
15. (2023). 2. Explainable Artificial Intelligence (XAI). Available from: 10.48047/ijfans/v12/i1/271
16. Sam, Sattarzadeh., Mahesh, Sudhakar., Konstantinos, N., Plataniotis., Jongseong, Jang., Yeonjeong, Jeong., Hyunwoo, Kim. (2021). 8. Integrated Grad-Cam: Sensitivity-Aware Visual Explanation of Deep Convolutional Networks Via Integrated Gradient-Based Scoring. Available from: 10.1109/ICASSP39728.2021.9415064
17. Erdem P. XAI Methods: Guided Grad-CAM [Internet]. erdem.pl. 2022 [cited 2025 Jan 1]. Available from: https://erdem.pl/2022/04/xai-methods-guided-grad-cam
18. Bodhisatwa, Mandal., Swarnendu, Ghosh., Teresa, Gonçalves., Paulo, Quaresma., Mita, Nasipuri., Nibaran, Das. (2021). 6. GuideBP: Guiding Backpropagation Through Weaker Pathways of Parallel Logits.. arXiv: Learning
19. Erfan, Hasanpour, Zaryabi., Loghman, Moradi., Bahareh, Kalantar., Naonori, Ueda., Alfian, Abdul, Halin. (2022). 1. Unboxing the Black Box of Attention Mechanisms in Remote Sensing Big Data Using XAI. Remote sensing, Available from: 10.3390/rs14246254
20. Transformer interpretability beyond attention visualization. Available at: https://www.researchgate.net/publication/355883670_Transformer_Interpretability_Beyond_Attention_Visualization
21. Jingxing, Zhang. (2024). 3. Classification and Comparison of Data Augmentation Techniques. Transactions on computer science and intelligent systems research, Available from: 10.62051/7e91md96
22. Perez, L. and Wang, J. (2017). *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1712.04621.
23. Xin, Jiang., Cheng, Xu., Zechao, Li. (2024). 1. Why pre-training is beneficial for downstream classification tasks?. Available from: 10.48550/arxiv.2410.08455
24. Ben, Graham., Alaaeldin, El-Nouby., Hugo, Touvron., Pierre, Stock., Armand, Joulin., Hervé, Jégou., Matthijs, Douze. (2021). 1. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. arXiv: Computer Vision and Pattern Recognition
25. Andrew, Howard., Andrey, Zhmoginov., Liang-Chieh, Chen., Mark, Sandler., Menglong, Zhu. (2018). 6. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation
26. Fengxiang, He., Tongliang, Liu., Dacheng, Tao. (2019). 6. Why ResNet Works? Residuals Generalize. arXiv: Machine Learning
27. (2023). 4. How Does Attention Work in Vision Transformers? A Visual Analytics Attempt. IEEE Transactions on Visualization and Computer Graphics, Available from: 10.1109/tvcg.2023.3261935
28. Zihang, Dai., Hanxiao, Liu., Quoc, V., Le., Mingxing, Tan. (2021). 3. CoAtNet: Marrying Convolution and Attention for All Data Sizes
29. Alaaeldin, El-Nouby., Hugo, Touvron., Mathilde, Caron., Piotr, Bojanowski., Matthijs, Douze., Armand, Joulin., Ivan, Laptev., Natalia, Neverova., Gabriel, Synnaeve., Jakob, Verbeek., Hervé, Jégou. (2021). 1. XCiT: Cross-Covariance Image Transformers.. arXiv: Computer Vision and Pattern Recognition,
30. Yao, Qiang., Chengyin, Li., Prashant, Khanduri., Dongxiao, Zhu. (2023). 1. Interpretability-Aware Vision Transformer. arXiv.org, Available from: 10.48550/arxiv.2309.08035