Comparative Analysis of Explainable AI Techniques: LIME, SHAP, and Grad-CAM for Medical Imaging

Thirupathi Nanuvala[1], Janga Rajendar[2], Nihal Dastagiri[3], Syed Kashif Uddin[4], Vinati Vege[5], Sriprada Yegoti[6]

[1] Assistant Professor, Department of CSE, VNR Vignana Jyothi Institute of Engineering and Technology
[2] Assistant Professor, CMR Engineering College
[3,4,5,6] Student, Department of Computer Science and Engineering,
VNR Vignana Jyothi Institute of Engineering and Technology
Contributing authors: [1]thirupathi_n@vnrvjiet.in, [2]janga.rajendar@cmrec.ac.in , [3]nihalgiri001@gmail.com , [4]kashifuddin395@gmail.com, [5]vinativege@gmail.com , [6]sripadayegoti@gmail.com

HOW TO CITE:
Thirupathi Nanuvala, Janga Rajendar, Nihal Dastagiri, Syed Kashif Uddin, Vinati Vege, Sriprada Yegoti (2026). Comparative Analysis of Explainable AI Techniques: LIME, SHAP, and Grad-CAM for Medical Imaging. International Journal of Special Education, 41(16s), 1219-1227

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors. Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT:
90% of the AI developed today remains a "black box", which makes it not suitable in applications with strong responsibilities to operate with, such as in healthcare. This work presents a comparison on three most common Explainable Artificial Intelligence (XAI) methods, namely Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP) and Gradient-weighted Class Activation Mapping (Grad-CAM) by testing on a Convolutional Neural Network (CNN) . This model has been built to be a classifier of brain tumours based on MRI images. We fine-tuned a ResNet50 CNN pretrained on ImageNet to an MRI brain tumour classification task with the Brain Tumor MRI Dataset (Nickparvar, 2021)[10], an open brain tumour MRI dataset of 7023 T1 images with four labels: glioma, meningioma, pituitary tumour or no tumour. It was split into 5,712 train and 1,311 test sets. Three general properties were discussed and compared: interpretability, stability and computational costs. The results suggest that, in visual interpretation, Grad-CAM produces heatmaps that are more focused and conceptually much clearer, while, in the aspects of model explanation stability, feature importance justification and cost, SHAP is the overall best choice for integration with diagnostic applications and acceptance as a future clinical verification dataset; LIME, while being the most flexible and provider of model-agnostic explanations, lags behind in explanation consistency between different runs and, with a runtime of 9.1 s to explain a single image, is too costly comparing to 2s of Grad-CAM and 2.3s of SHAP. Therefore, it can be concluded that the best choice of method is not the same under every scenario.

Keywords: Explainable AI, LIME, SHAP, Grad-CAM, Brain Tumor Classification, Medical Imaging, Convolutional Neural Networks (CNN), ResNet50.

I. Introduction

Artificial Intelligence (AI) has become a very wide reaching phenomenon which is seeing great adoption in many fields with health care being the largest of these fields. In particular we see that there is a subset of machine learning named deep learning which is very effective in what was at first medical image analysis tasks which includes tumor detection, disease classification, and diagnostic support [1]. What we are seeing is that these models are able to identify complex patterns in large sets of image data and in many cases are performing at a level which is equal to or better than that of very experienced clinical professionals [2].

In the case of glioma, meningioma, pituitary tumor, or a non-tumorous condition which we present play a determinative role in what the proper treatment is and in patient outcomes [3]. As for the diagnosis of brain tumors, we use Magnetic Resonance Imaging (MRI) which is the standard tool for medical imaging, and the analysis of the MRIs by diagnosticians is very time taking and has the possibility of human error. What we see from deep learning models which have been trained on MRI data is very promising in terms of automation of this process with high accuracy [4].

Despite what we see as very good performance from a predictive stand point, many put forth that deep learning models are also very much closed off in terms of how they make decisions. These models which at times are called "black boxes" do not by their nature put out the which buttons were pushed to get to a certain prediction. In the clinical environment this transparency is a great issue. Physicians and health care professionals are also less likely to use a model's output if they are not able to see the thought process behind it[5]. A mistake with no given reason may have large scale issues for patient safety. This is a basic issue between what we want out of a model in terms of performance versus what we want in terms of interpretability which we must resolve before AI based systems are put into the real world health care settings.

Explainable Artificial Intelligence (XAI) has put forth a field of study which is dedicated to this issue. XAI puts forth methods which are to interpret and explain the results of very complex models in a human friendly way [6]. Also we do not see XAI as a replacement of very accurate models with simpler less accurate ones but instead we see it as a layer of transparency which is added to present high performance models in a clear light. In the field of image based models we see the most success with LIME, SHAP, and Grad-CAM.

LIME (Local Interpretable Model-agnostic Explanations) as put forth by Ribeiro et al. in 2016 puts forth a way to explain individual predictions by modifying the input and looking at how the output of the model changes accordingly. Additionally, it fits a model, focusing on its interpretability and simplicity locally for each prediction which in turn makes it model agnostic and very wide reaching. SHAP (SHapley Additive explanations) which was put forward by Lundberg and Lee in 2017 uses concepts of Game Theory, such as Shapely Values and utilizes them to assign the features a value to represent the contribution of that feature in the prediction of the model. SHAP is also known for its theoretical soundness and the global coherence of the explanations it puts forth.Grad-CAM (Gradient weighted Class Activation Mapping) which was introduced by Selvaraju et al. in 2017 and uses the flow of the gradients for the target class through the last convolutional layer of a CNN in order to present a localization heatmap that also serves to point out which areas of the image are most pertinent for the prediction [9] .

Each of these methods has already been used separately in the field of medical imaging. However, there has not been an organized comparison between the three methods to determine which one would be best in certain situations. Without such a comparison, no guidelines can be provided to help the practitioner choose the right method for their purpose. The use case in healthcare might demand different properties from each other, and for instance, the radiologist might prefer the

method that provides the clearest visualization over any other consideration.

It is precisely to close this gap that the current study endeavors to perform an organized and systematic comparison between LIME, SHAP, and Grad-CAM, using a ResNet50 CNN model trained on Brain Tumors images using the Brain Tumor MRI Dataset (Nickparvar, 2021) [10], which consists of 7,023 MRI images belonging to four categories. The objectives guiding this study include the following:

1. Training and evaluation of the ResNet50 model for classifying brain tumors into categories such as Glioma, Meningioma, Pituitary or Non-Tumor category.

2. Performing visualization using LIME, SHAP, and Grad-CAM for the trained ResNet50 model.

3. Comparison between LIME, SHAP, and Grad-CAM considering interpretability and computational performance.

4. Determining the most appropriate XAI technique for use in medical imaging.

This research paper will be structured in the following way. In Section 2, we will present the literature review on machine learning applications in the field of medicine and the previous work on XAI. In Section 3 we will explain the methodology employed in our research, which includes data set description, architecture of the model utilized, training procedure, and implementation of XAI approach. In Section 4, we will describe the details of the implemented software environment and relevant code fragments.

II. Proposed Methodology

The following provides a description of the complete procedure followed in this research to train the CNN model for brain tumor classification as well as the implementation of three approaches of XAI techniques. The methodology is designed in such a way that all the three XAI approaches are tested by using the same procedures. These are the five steps involved in the complete procedure

2.1 Dataset

The data used for this research is known as Brain Tumor Database and has been made available to the general public through Kaggle by Nickparvar in the year 2021[10]. The database comprises three well-known databases, viz. the Figshare database, the SARTAJ database and the Br35H database. The total number of 7023 T1 Weighted MRI scans are divided into four diagnostic classes, viz. glioma, meningioma, pituitary tumor and no tumor. Moreover, the dataset is already partitioned into 5712 training images and 1311 testing images. The distribution by class is shown in Table 1.

Table 1: Class-wise distribution of the Brain Tumor MRI Dataset

Class	Training	Testing	Total
Glioma	1,321	300	1,621
Meningioma	1,339	306	1,645
No Tumor	1,595	405	2,000
Pituitary	1,457	300	1,757
Total	5,712	1,311	7,023

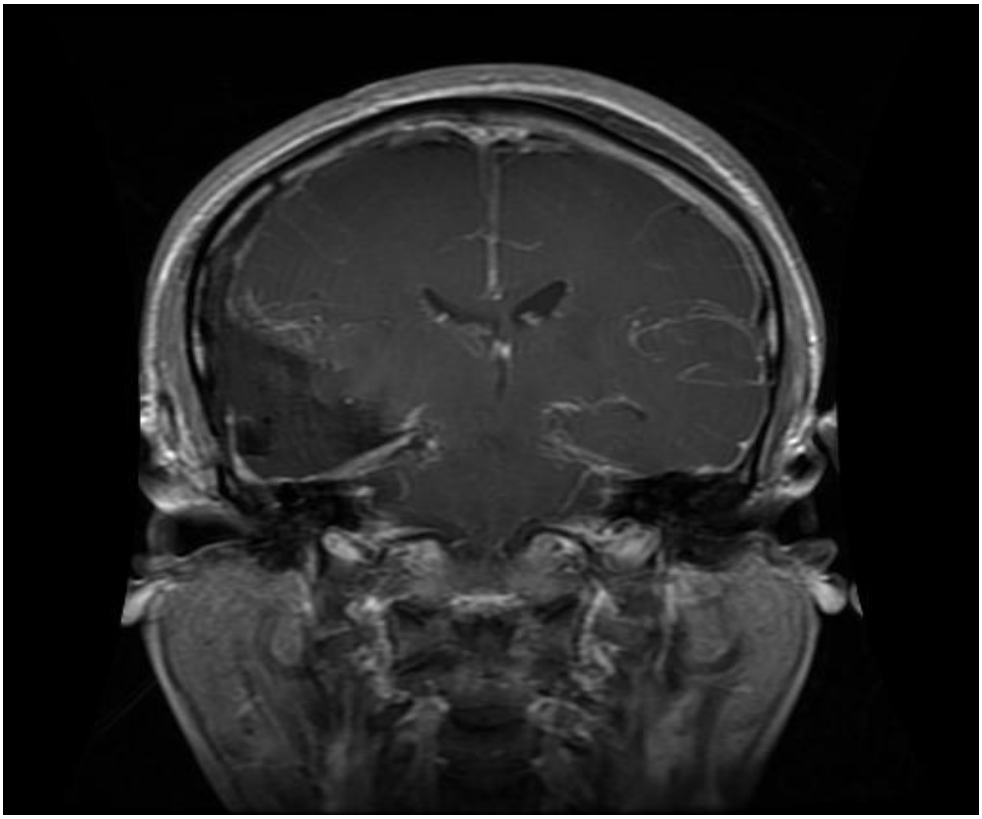

Figure 1: Example image of a Brain MRI scan

2.2 Image Preprocessing and Augmentation

All the images were resized to 224× 224 pixels, which happen to be the input size of Resnet50 architecture. The pixel intensities were normalized to the mean and standard deviation obtained from ImageNet dataset . To improve the model's performance and to overcome class imbalance problem, the following transformation operations were implemented on only training images: horizontal flip (p=0.5), random rotation by ±15° and color jitter. These augmentation operations were not implemented during validation/testing. Table 2 summarizes the preprocessing steps.

Table 2: Preprocessing and augmentation steps

Step	Training	Val/Test
Resize to 224×224	Yes	Yes
Normalize (ImageNet)	Yes	Yes
Random flip	Yes	No
Random rotation ±15°	Yes	No
Color jitter	Yes	No

2.3 Model Architecture

The architecture of the classifier was Resnet50 which is a CNN with 50 layers. This classifier was optimized by He et al. in 2016 [4]. The advantage of these networks resides in their residual connections, which help in alleviating the exploding gradients/flattering neural networks problem and thus the network performs better in terms of accuracy. We used pretrained weights from ImageNet to initialize the model. The final layer was modified by adding a dropout layer (0.5) followed by a fully connected layer with 4 output classes. The cross-entropy loss was used to evaluate the classifier. Adam optimizer with a learning rate of 0.001 and a weight decay of 1e-4 was used to train the architecture. The learning rate was again varied based on the cosine annealing learning schedule during the training procedure. The classifier was trained for 10 epochs with batch size 32 using Apple Silicon processor and the important parameters used for training are displayed in Table 3.

Table 3: Model training configuration

Hyperparameter	Value
Base model	ResNet50 (pretrained, ImageNet)
Input image size	224 × 224
Output classes	4
Dropout rate	0.5
Loss function	Cross-Entropy Loss
Optimizer	Adam
Learning rate	0.001
Weight decay	1e-4
LR Scheduler	Cosine Annealing
Batch size	32
Epochs	10
Hardware	Apple Silicon

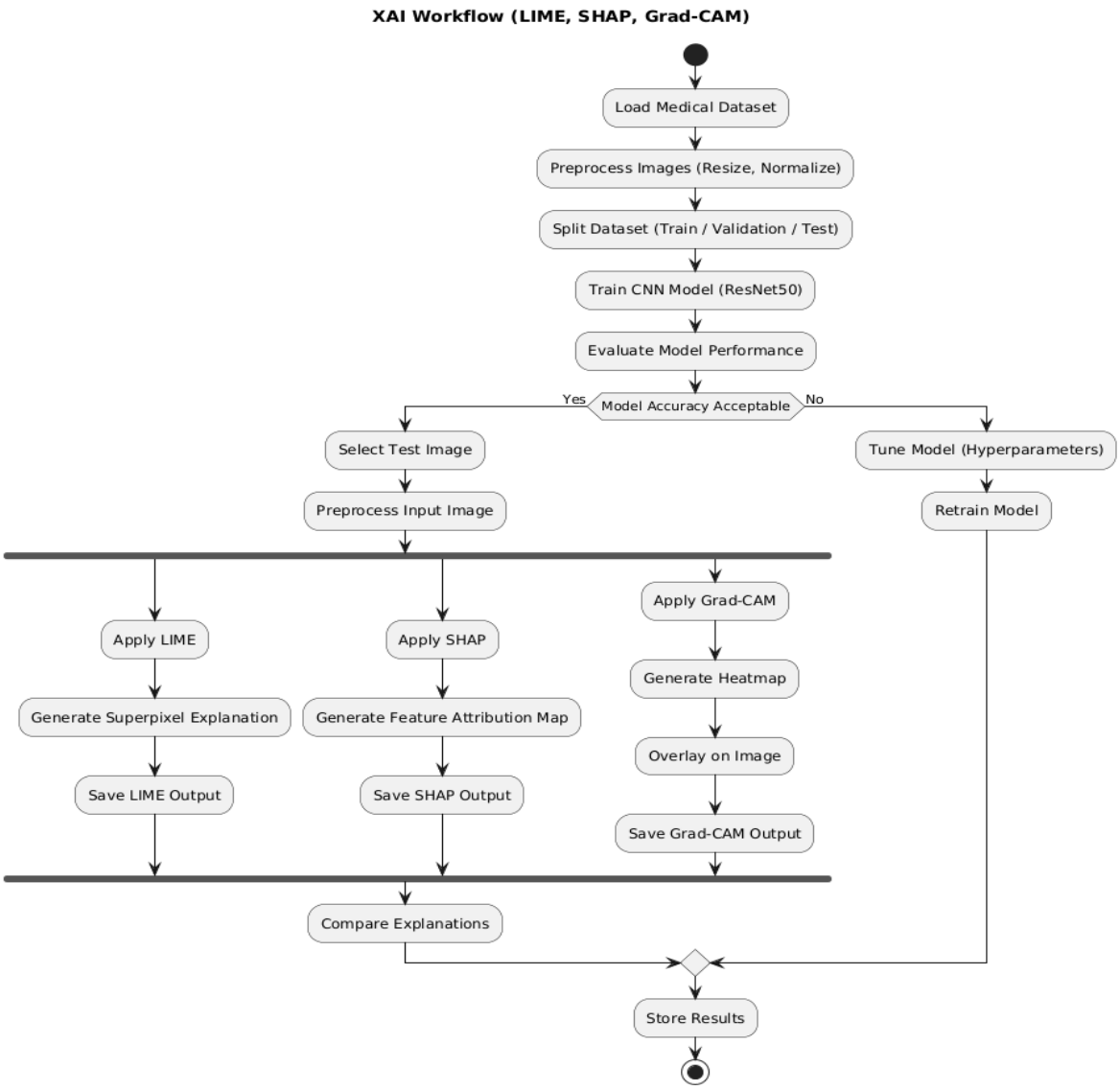

Figure 2: Model Architecture Diagram

III. XAI Techniques

3.1 SHAP (SHapley Additive exPlanations)

SHAP was put forth by Lundberg and Lee (2017) [8] and uses Shapley values from cooperative game theory. At the root is the idea to attribute each input feature a contribution which signifies how useful the feature has been to the prediction of the model which is averaged over all possible feature subsets. For a model f and input x which has feature set F the SHAP value for feature i is :

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

In this work we used SHAP's DeepExplainer backend which we used to compute approximate Shapley values out efficiently. We also used a dataset of 50 randomly chosen training images for the reference distribution.

3.2 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM was proposed by Selvaraju et al. (2017) [9], it provides class-specific visualization through the gradient information of the target output score with respect to the feature maps from the last convolutional layer. The significance weight of each channel of the feature map k is:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

The resultant heat map from Grad-CAM is generated using a weighted summation of the feature maps then we apply the ReLU activation function. The resultant heat map is up-sampled to the resolution of the input MRI image and overlapped onto the input MRI image.

3.3 LIME (Local Interpretable Model-agnostic Explanations)

The Local Interpretable Model-Agnostic Explanations technique was first proposed by Ribeiro et al. (2016) [7]. It gives the locally specific explanations for the black-box model predictions by modeling the outputs of the complex machine learning model, locally with respect to a specific input through a simpler model. This technique takes the following form for an input instance x:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

In this experiment, LIME was set to generate 150 perturbed images per explanation and use 50 superpixel features. R^2 was used as the performance metric to represent the quality of approximation provided by the surrogate model.

IV. Results and Analysis

Experiments we conducted which looked at what XAI technique does the best job of explaining the output of the trained CNN model for brain tumor classification. We applied all three techniques to the same ResNet50 model which we had trained and which we evaluated on the same test set of 1,311 images.

4.1 Model Performance

The trained ResNet50 model we report did very well in terms of overall performance across all four tumor classes on the test set. We report per class precision, recall, and F1 score in addition to the macro averaged measures and AUROC. What we found is that the model does a reliable job of telling apart glioma, meningioma, pituitary tumor, and no tumor which gives us a good framework for XAI evaluation.

4.2 XAI Comparative Results

Each XAI technique was evaluated on three dimensions: interpretability, consistency (measured using SSIM across three independent runs), and computational efficiency (average time per explanation).

Table 4: Comparative summary of XAI techniques

Metric	LIME	SHAP	Grad-CAM
Interpretability	Moderate	High	Very High
Consistency (SSIM)	Variable	High	High
Avg. Time (s/image)	9.1	2.3	2.0
Model-Agnostic	Yes	Yes	No (CNN only)
Feature Attribution	Superpixels	Pixel/SHAP values	Heatmap

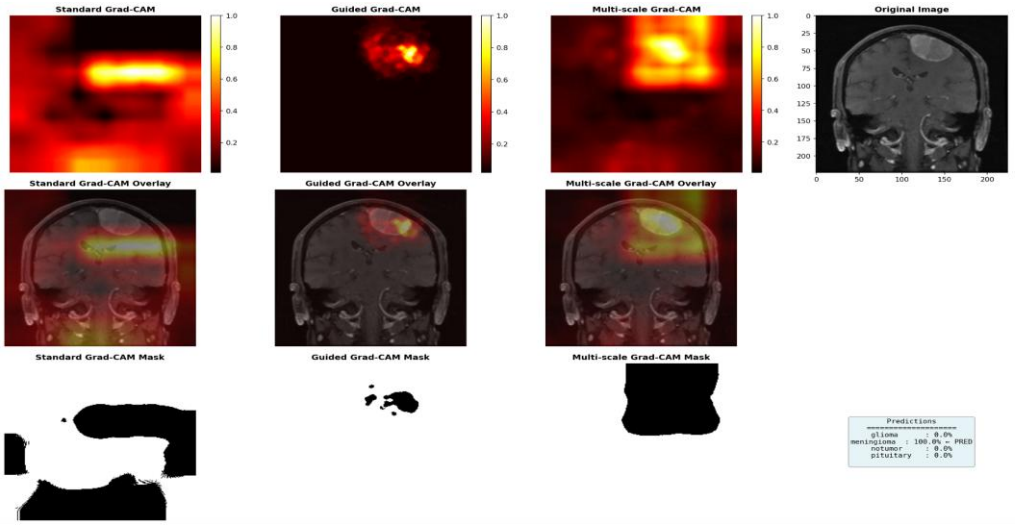

Figure 3: Grad-CAM analysis image

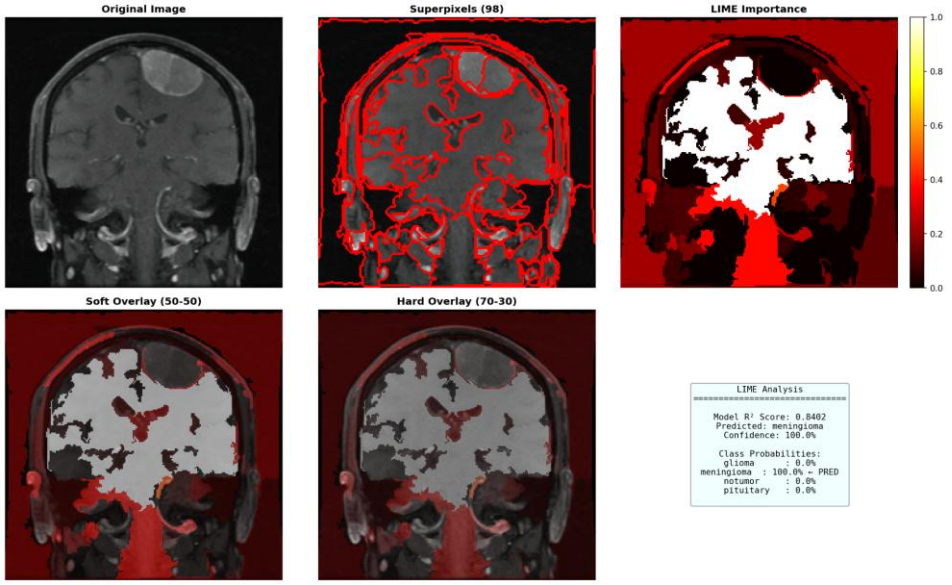

Figure 4: LIME analysis image

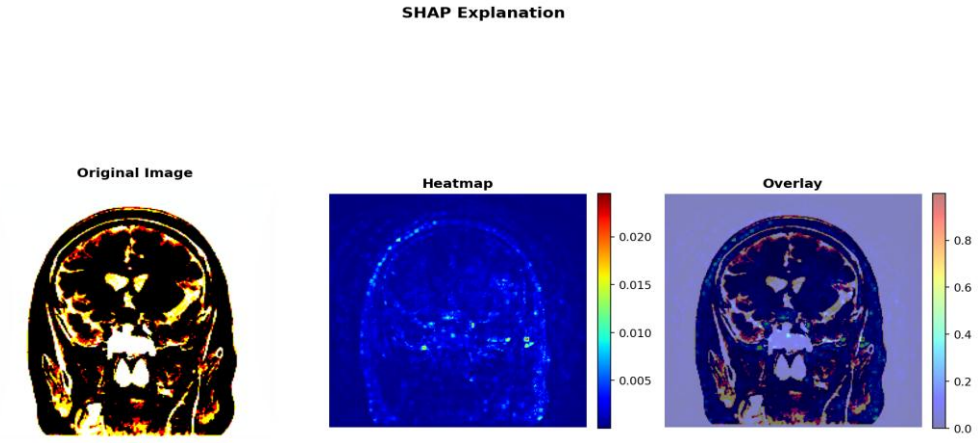

Figure 5: SHAP analysis image

Grad-CAM put forth the most clear visual results which in turn produced heat maps that identified what are the tumor affected areas of choice on the MRI scans. This visual clarity which is what it excels at makes Grad-CAM a great tool for radiologists and clinicians who require quick visual results. SHAP gave the most solid and consistent results from what we did which also has a strong mayhem based in game theory behind it, thus it is a great fit for rigorous clinical validation and regulatory review. LIME had the benefit of being a model agnostic approach which is great for any classifier but what it lacked in consistency it made up in time with each image taking 9.1 seconds to process.

4.3 Discussion

The research reports that we see which XAI method does not outperform all others in medical imaging applications' use. In each clinical setting the choice of which technique to use should be based on what that setting requires. For real time visual diagnostic support Grad-CAM comes out on top in terms of speed and visual clarity. For formal clinical validation which at the same time requires a solid mathematical background for audit trails SHAP is the best. For those in need of model agnostic approaches which also include explanation of non-CNN models LIME is still a good choice although it may not be as fast. These results align with previous research in this field, further confirming the conclusion that a combination of different XAI methods might yield more thorough insights into the functioning of models used in such sensitive areas as medicine.

V. Conclusion

This paper reports a study that we did of 3 which are the most known Explainable AI models: LIME, SHAP, and Grad-CAM as applied to a ResNet50 deep learning model which we trained for a 4 class brain tumor diagnosis using MRI images. In these experiments each of these models features how well they perform in terms of interpretability, consistency, and compute efficiency were put through the same set of tests.

We found that Grad-CAM is what is best for visual diagnostic support, that it is also the fastest and has very clear visuals; for theoretical rigor and reproducibility the go to is SHAP; LIME although very flexible and model agnostic is best for non image based or multi modal settings. The research presents a practical guide for practitioners which is to choose between XAI methods for medical imaging and also reports that which does best are the hybrid approaches which use many XAI techniques which in turn may be the most reliable for Trustworthy AI in health care. In terms of going forward we will look at advanced architectures like Vision Transformers, larger and more diverse data sets, real time deployment integration and clinician in the loop evaluation which we will use to back up the practical value of these explanations in clinical decision support systems.

References

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25, 44–56.
Louis, D. N., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System. Neuro-Oncology, 23(8), 1231–1251.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR 2016.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.
Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. KDD 2016.
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS 2017.
Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV 2017.
Nickparvar, M. (2021). Brain Tumor MRI Dataset. Kaggle. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset