Nimish Gupta

Explainable AI for Brain Tumor Detection using GradCAM, Smooth GradCAM and LIME

Dhruv Aggarwal, Nimish Gupta, Ananya Choudhary, Dr. Abhinav Tomar, Dr. Vijay Kumar Bohat, Dr. Gaurav Singal

Netaji Subhas University of Technology, Dwarka, Delhi, 110078, India

Abstract

Since the start of the current century, artificial intelligence has gone through critical advances improving the capabilities of intelligent systems. Especially machine learning has changed remarkably and caused the rise of deep learning. Deep learning shows cutting-edge results in terms of even the most advanced, difficult problems. However, that includes a trade-off in terms of interpretability. Although traditional machine learning techniques employ interpretable working mechanisms, hybrid systems and deep learning models are black-box being beyond our understanding capabilities. So, the need for making such systems understandable, additional methods by explainable artificial intelligence (XAI) has been widely developed in last years. In this sense, this study purposes a Convolutional Neural Networks (CNN) model, which is explainded by various xai algorithms such as Grad-CAM, smooth GradCAM and LIME. As providing numerical feedback in addition to the default Grad-CAM, the smooth GradCAM was used within the developed CNN model, in order to have an explainability interface for brain tumor diagnosis. In detail model was evaluated via technical and physicians-oriented (human-side) evaluations. The model provided average findings of 95.00 percent of train accuracy and 99.85 percent of train accuracy.

Keywords: XAI , explainability , interpretability , GradCAM , LIME , Smooth GradCAM

1. Introduction

The abnormal and uncontrolled growth of cells anywhere in the body is termed cancer. There are approximately more than 200 types of cancer including, lung cancer, breast cancer, skin cancer, blood cancer, heart cancer, and lymphoma. According to the World Health Organization (WHO), cancer is the second leading cause of death in the world with around 9.6 million deaths in 2018.

Brain tumors, a common and aggressive disease are the cause of the highest rate of short life expectancy amongst many people of different age and gender groups. A brain tumor is a mass of abnormal cells in the brain. It usually occurs when there is an abrupt and unusual extension of tissues in the brain. The number of these abnormal cells does not remain constant instead, they are multiplied very rapidly and begin to spread themselves. A brain tumor can be

cancerous (malignant) or non-cancerous (benign). The malignant tumors can quickly spread to other tissues in the brain and it might worsen the patient's condition. Normal brain cells are replaced by new cells when they become old or damaged. If already present old and damaged cells are not eliminated fast, it can cause a huge problem by forming an extra mass of unnecessary brain tissue, thus leading to a tumor. Because a tumor normally spreads itself to the neighboring tissues, therefore, if the tumor is diagnosed and treated early, the chances of effective treatment and hence survival become high.

Medical imaging plays a very vital role in the diagnosis of brain tumors. With the help of MRI and Computed Tomography (CT) scans, necessary information about the presence and spread of abnormal tissues in the brain is provided so that the status of the tumor could be known and necessary follow-up can be made on time. MRI is preferred over CT scans because CT scans use radiation like X-rays to generate images of internal body parts while MRI does not use harmful radiation for imaging purposes. MRI uses radio waves and a powerful magnetic field and has the capability of providing detailed information about the internal structure of the human body and soft tissues. MRI generates Proton Density (PD), FLAIR, T1, T1 contrast-enhanced, and T2 weighted high contrast grayscale images.

A brain tumor is categorized and differentiated based on size, shape, and location in the brain. A brain tumor can be either primary or secondary, where the primary tumor is originated from within the brain cells and the secondary tumor spreads itself to the brain from another part of the body. Meningioma is the most common benign primary tumor. Gliomas on the other hand are the most pervasive type of adult brain tumor responsible for 78 percent of malignant tumors. WHO has categorized tumor types into grades concerning growth rate, the ratio of malignancy, recurrence, and aggressiveness of the tumor starting from grade 1 to IV. Grade I and II are low-grade tumors whereas grades III and IV are high-grade tumors.

Today's modern, brave world is rising over many advanced technological advances. As a result of different industrial phases triggering transformation of both technological tools and the societies, the current world has reached to the era of intelligent systems. It is remarkable that the field of artificial intelligence passed through many different development processes since its first introduction to the scientific audience. Although there are different methods and techniques used in the context of artificial intelligence field, machine learning is known as the main actor on the background of successful applications . Machine learning algorithms, which are essentially based on the logic of providing adaptation to the target training data by optimizing various parameters inside the model, have been affected by intense advances within hardware and software technologies, seen in especially last twenty years. More effective techniques having better capabilities to process more challenging data are now accepted under the name of new machine learning: deep learning . For now, neural network-oriented techniques build the main characteristics of deep learning. It is clear that the future has the potential of newer deep learning models out of the neural network formation but the deep learning has been shaped with advanced neural network models as a result of the needs of different problem areas and alternative solution mechanisms. Here, Convolutional Neural Networks (CNN) has a remarkable popularity among all deep learning techniques with its successful applications in different fields. Various CNN models have been used effectively in image-based problems, especially thanks to integrated feature extraction mechanisms and layer structures designed primarily based on image data. Because of that, CNN

models are often used in critical areas such as healthcare, which have been closely linked with artificial intelligence through the historical advances. Among many other healthcare-oriented problem areas, disease diagnosis is widely employing CNN models to get successful outcomes. CNN models are often reported as they provide more sensitive, early detections when compared to physicians.

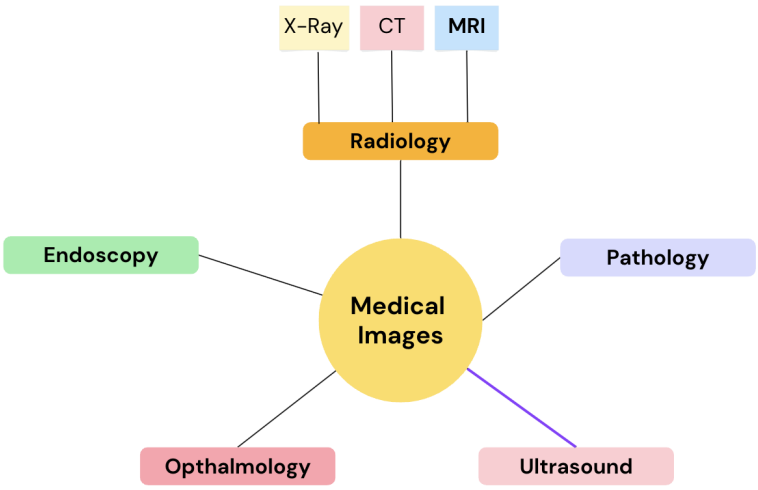

Figure 1: Applications of Medical Images in different fields of healthcare

Nowadays, CNN is among top intelligent tools to deal with disease diagnosis problems. Although there are alternative deep learning models, CNN has already shown its critical role in especially medical image-based diagnosis problems. It seems that CNN and deep learning era takes the outcomes by hybrid machine learning formations some more steps away. However, when examined under the human-side understanding capabilities, it is often discussed that the solution mechanisms by both hybrid machine learning formations and the recent deep learning models are black-box, meaning that we cannot control well enough the decision-making mechanism between input data and the outcomes. When we consider the CNN model, the total number of parameters reaching to hundreds and even thousands cause us to lose our tracking control among different mathematical calculations. At this point, it is not possible to understand safety level of the decisions made by the CNN-based intelligent system. Furthermore, it is also not possible to catch any errors as well as bias or understand success and fail borders of the used system. This situation affecting not only present state of the artificial intelligence field but also future advances of the upcoming intelligent systems has enabled researchers to search for effective solutions. Nowadays, these solutions include using integration of interpretable machine learning techniques and building new mathematical components to make black-box models transparent (or white-box). The efforts so far caused a new research area: explainable artificial intelligence (XAI) to rise, and that area has been effective for designing alternative models of deep learning techniques, building explainability interfaces for medical diagnosis applications.

Based on the explanations so far, the main objective of this study is to purpose an explainable CNN for brain tumor diagnosis problem. As it is known, CNN models have been often used in diagnosis from medical image data including X-Ray, CT, MRI, ultrasound. . . etc.. For supporting explainability of the verbal decision-makings from such medical images, Class Activation Mapping (CAM) method is used for creating heat maps over input medical image and showing which regions are detected by the model during detection (diagnosis). The CAM includes different variations currently and the Grad-CAM is among these. However, according to the authors, the Grad-CAM is still open for further updates. So, this study aimed to improve it and use with the CNN, for a remarkable diagnosis problem. The target problem was chosen as brain tumor diagnosis, as it has gained momentum in recent years. Pointing the main objective, the motivations of the study can be expressed briefly as follows:

1. Building a successful enough CNN model for brain tumor diagnosis over a dataset from the literature.
2. Ensuring XAI touch for a CNN model solving brain tumor diagnosis.
3. To use GradCAM , Smooth GradCAM and LIME to detect the most significant regions in an image.

Future work - Performing critical evaluations with the physicians to see if the developed GradCAM-CNN model is successful and safe enough for the brain tumor diagnosis problem.

2. Literature Survey

2.1. Tumor Segmentation from MRI Brain Images: Challenges and Progress

One of the most challenging as well as demanding task is to segment the region of interest from an object and segmenting the tumor from an MRI Brain image is an ambitious one. Researchers around the world are working on this field to get the best-segmented ROI and various disparate approaches simulated from a distinct perspective.

2.2. Recent Advances in Brain Tumor Segmentation: Neural Network-Based Approaches

Nowadays Neural Network based segmentation gives prominent outcomes, and the flow of employing this model is augmenting day by day. Devkota et al established the whole segmentation process based on Mathematical Morphological Operations and spatial FCM algorithm which improves the computation time, but the proposed solution has not been tested up to the evaluation stage and outcomes as- Detects cancer with 92 and classifier has an accuracy of 86.6 percent. Yantao et al resembled Histogram based segmentation technique. Regarding the brain tumor segmentation task as a three-class (tumor including necrosis and tumor, edema and normal tissue) classification problem regarding two modalities FLAIR and T1. The abnormal regions were detected by using a region-based active contour model on FLAIR modality. The edema and tumor tissues were distinguished in the abnormal regions based on the contrast enhancement T1 modality by the k-means method and accomplished a Dice coefficient and sensitivity of 73.6 percent and 90.3 percent respectively. Based on edge detection approaches, Badran et al adopted the canny edge detection model accumulated with Adaptive thresholding to extract the ROI.

2.3. Improved Brain Tumor Detection with Canny Edge Detection vs. Adaptive Thresholding

The dataset contained 102 images. Images were first preprocessed, then for two sets of a neural network, for the first set canny edge detection was applied, and for the second set, adaptive thresholding was applied. The segmented image is then represented by a level number and characteristics features are extracted by the Harris method. Then two neural networks are employed, first for the detection of healthy or tumor containing brain and the second one is for detecting tumor type. Depicting the outcomes and comparing these two models, the canny edge detection method showed better results in terms of accuracy.

2.4. Novel Approach for MRI-Based Tumor Segmentation Using Growth Patterns

Pei et al proposed a technique which utilizes tumor growth patterns as novel features to improve texture based tumor segmentation in longitudinal MRI. Label maps are being used to obtain tumor growth modeling and predicting cell density after extracting textures (e.g., fractal, and mBm) and intensity features.

2.5. Improved MRI Tumor Segmentation with Probabilistic Neural Networks

Performance of the model reflected as the Mean DSC with tumor cell density- LOO: 0.819302 and 3-Folder: 0.82122. Dina et al introduced a model based on the Probabilistic Neural Network model related to Learning Vector Quantization. The model was evaluated on 64 MRI images, among which 18 MRI images were used as the test set, and the rest was used as a training set. The Gaussian filter smoothened the images. 79 percent of the processing time was reduced by the modified PNN method. A Probabilistic Neural Network based segmentation technique implemented by Othman et al. Principal Component Analysis (PCA) was used for feature extraction and also to reduce the large dimensionality of the data . The MRI images are converted into matrices, and then Probabilistic Neural Network is used for classification. Finally, performance analysis is done. The training dataset contained 20 subjects, and the test dataset included 15 subjects. Based on the spread value, accuracy ranged from 73 percent to 100 percent.

2.6. Understanding Deep Model Explainability: Instance vs. Model-level Explanations

Explainability based on deep models can be divided into input-dependent explanations (instance-level explanations) and input-independent explanations (model-level explanations). The former can find and explain the features that have the greatest impact on the prediction results, while the latter can directly explain the model without considering the network input.

2.7. Explaining Medical Imaging: Perturbation vs. Gradient Methods

In medical imaging, with the clarity of segmentation and classification, explainability of these tasks usually adopt instance-level explanations, of which the more widely used are perturbation-based methods and gradient-based methods. In the early stage of medical image explainability work, perturbation-based methods are often used. These methods can study the network by observing output changes under different input disturbances. In the medical field, disturbances can include various forms, such as shape and occlusion. However, since the running time depends on the number of input features, more computing resources and time are often needed to achieve better results, so more consideration should be given to the choice between accuracy and explainability.

2.8. Insights into Gradient-Based XAI Methods

In gradient-based methods, gradient is the approximate value of importance of the input feature and is highly related to model parameters. Similar models often have the advantage of post-processing. Explainability can be independent of the model training process, avoiding the balance between accuracy and computational loss. Gradients used in XAI include integral gradient (IG) , vanilla gradient (VG), and guided backpropagation (GB).

2.9. Advances in XAI for Medical Imaging Tasks

In recent years, researchers have used XAI to comprehensively evaluate and explain model results. Natekar et al used Grad-CAM to explain the brain tumor segmentation task. Adebayo et al found that CAM-based methods are better in classification tasks through the sanity check. Pereira et al proposed an explainable method combining global and local information for tumor segmentation. They used GB and CAM in brain tumor detection. The experimental results show that GB can distinguish important areas rather than categories, and CAM works well in both tasks. Narayanan et al used GoogLeNet and ResNet to detect malaria, diabetic retinopathy, brain tumors, and tuberculosis in different imaging modalities. They visualized the class activation mappings to enhance the understanding of these deep networks.

2.10. Explainability in Medical Imaging: Building Trust

The explainability of medical images is important and has great demand and scalability because it is related to the degree of trust of medical professionals in model results and subsequent operations

3. Comparison

In the world of medicine, explainable artificial intelligence (XAI) is becoming increasingly important because it has the potential to completely change the way we do healthcare.
Firstly, XAI helps to minimize the chances of errors in medical diagnoses by reducing the subjectivity that can sometimes creep into the process. By providing clear explanations for why it makes certain diagnoses, XAI systems help doctors understand the logic behind AI-generated diagnoses, which in turn helps them make more confident decisions.
But it doesn't stop there. When we bring XAI into the realm of medical imaging, we open up a whole new world of possibilities for personalized medicine. By showing doctors the underlying patterns and features that AI models use to make predictions, XAI allows for treatments and interventions to be tailored specifically to each patient's needs. This not only makes treatments more effective but also helps minimize side effects and ensures that resources are used efficiently.
Another big benefit of XAI is that it promotes fairness and accountability in healthcare. By making AI algorithms transparent and easy to understand, XAI helps identify and address any biases that might affect medical diagnoses and treatments. This builds trust between patients and doctors and ensures that everyone has access to high-quality healthcare, regardless of who they are or where they come from.
And let's not forget about ethics. By sticking to principles like transparency, accountability, and putting patients first, XAI-driven healthcare systems ensure that patient rights are protected and that ethical standards are upheld in medical research and practice.
In summary, XAI is a game-changer in the world of medicine. It improves diagnostic accuracy, makes personalized medicine possible, ensures fairness and accountability, and upholds ethical

standards. By choosing XAI over blackbox methods, we're not just improving healthcare – we're transforming it for the better.

4. Components and Architecture Used

4.1. Dataset Overview

The dataset used in this study consists of images obtained from medical imaging archives and repositories. These images are categorized into four classes based on the presence or absence of specific types of brain tumors: meningioma tumor, no tumor, pituitary tumor, and glioma tumor. The dataset is taken from this GitHub repository.

The distribution of images in each class is as follows:

- **Meningioma Tumor**: This class comprises 937 images depicting cases where meningioma tumors are present.
- **No Tumor**: This class includes 501 images representing brain scans with no detectable tumors or abnormalities.
- **Pituitary Tumor**: With 901 images, this class encompasses cases with pituitary tumors, a common type of brain tumor originating from the pituitary gland.
- **Glioma Tumor**: The dataset contains 926 images depicting cases of glioma tumors, which are malignant tumors that arise from glial cells in the brain.

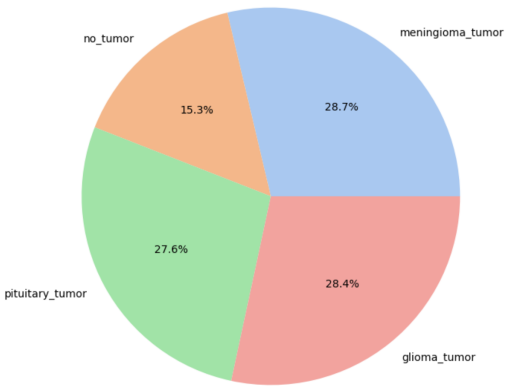

Figure 2: Distribution of images in dataset

The selection of these classes provides a comprehensive representation of various pathological conditions encountered in neuroimaging. The dataset enables the development and evaluation of machine learning models for brain tumor classification, facilitating advancements in medical image analysis and diagnosis.

Additionally, the dataset undergoes preprocessing and augmentation techniques to enhance its diversity and balance, ensuring robust model training and evaluation. These steps include normalization, resizing, and data augmentation methods such as rotation, shifting, and flipping.

The utilization of a well-curated and diverse dataset is crucial for the development and evaluation of accurate and reliable machine learning models for medical image analysis. It enables researchers and practitioners to address real-world challenges in healthcare and contribute to improved diagnostic and therapeutic outcomes for patients with brain tumors.

4.2. Architecture: EfficientNetB1

The EfficientNetB1 architecture is a part of the EfficientNet family of convolutional neural networks (CNNs) proposed by Tan et al. (2019). It is characterized by its efficient use of parameters and computational resources, making it suitable for deployment in resource-constrained environments such as mobile devices and embedded systems.

EfficientNetB1 leverages a novel compound scaling method that uniformly scales network width, depth, and resolution to achieve optimal performance. This compound scaling approach ensures that the network is well-balanced across different architectural dimensions, allowing it to achieve better accuracy with fewer parameters compared to traditional CNN architectures.

At the core of EfficientNetB1's architecture are depthwise separable convolutions, which decompose the standard convolution operation into separate depthwise and pointwise convolution layers. This decomposition reduces computational complexity while preserving representational capacity, resulting in more efficient and lightweight models. Another key component of Effi-

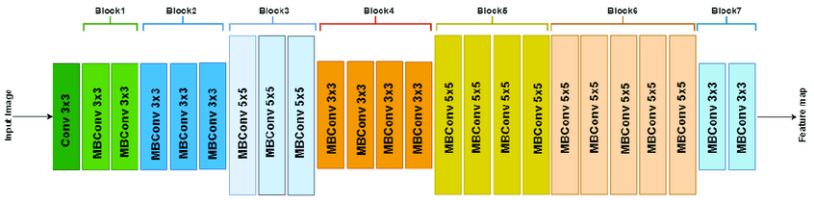

Figure 3: Distribution of images in dataset

cientNetB1 is the squeeze-and-excitation (SE) block, which enhances feature representation by adaptively recalibrating channel-wise feature responses. The SE block comprises two operations: squeeze, which globally pools feature maps to capture channel-wise statistics, and excitation, which learns channel-wise importance weights to amplify informative features.

EfficientNetB1 also incorporates a compound scaling method that balances model size and computational cost across different architectural dimensions. This method systematically scales network width, depth, and resolution based on a compound coefficient, allowing for efficient model design without compromising performance. In addition to its architectural components, EfficientNetB1 benefits from pretraining on large-scale image datasets such as ImageNet. Pretraining enables the model to learn generic image representations that capture high-level semantic features, which can then be fine-tuned on domain-specific tasks with limited labeled data.

In this study, EfficientNetB1 serves as the backbone architecture for brain tumor classification, leveraging its efficient design and pretrained weights to achieve accurate and efficient classification of brain tumor images. By harnessing the power of EfficientNetB1, we aim to improve the diagnosis and treatment of brain tumors through automated image analysis and interpretation.

4.3. XAI (Explainable Artificial Intelligence)

4.3.1. GradCAM (Gradient-weighted Class Activation Mapping)

Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. See Figure 4 for reference.

The technique is an improvement over previous approaches in versatility and accuracy. It is complex but, luckily, the output is intuitive. From a high-level, we take an image as input and create a model that is cut off at the layer for which we want to create a Grad-CAM heat-map. We attach the fully-connected layers for prediction. We then run the input through the model, grab the layer output, and loss. Next, we find the gradient of the output of our desired model layer w.r.t. the model loss. From there, we take sections of the gradient which contribute to the prediction, reduce, resize, and rescale so that the heat-map can be overlaid with the original image.

4.3.2. SmoothGradCAM (Smooth Gradient-weighted Class Activation Mapping)

Smooth Grad-CAM tackles the lack of transparency by offering visual explanations. It builds on Grad-CAM, which highlights influential image regions by analyzing gradients. However, Grad-CAM can be noisy. Smooth Grad-CAM addresses this by computing gradients over multiple noisy versions of the input image, resulting in smoother explanations. This leads to sharper visualizations and better object localization. Smooth Grad-CAM is a valuable tool for researchers to understand how CNN models make predictions, promoting explainability in deep learning.
For each noisy version, it analyzes the gradients associated with the image prediction. By averaging these analyses, it creates a smoother "importance map" highlighting the image areas truly critical for the model's decision. This map helps researchers visualize which visual cues, like the extra mass in brain's MRI image, were most influential, promoting explainability and trust in deep learning applications.

4.3.3. LIME (Local Interpretable Model-agnostic Explanations)

The beauty of LIME its accessibility and simplicity. The core idea behind LIME though exhaustive is really intuitive and simple! Let's dive in and see what the name itself represents:

- Model agnosticism refers to the property of LIME using which it can give explanations for any given supervised learning model by treating it as a 'black box' separately. This means that LIME can handle almost any model that exists out there in the wild!

- Local explanations mean that LIME gives explanations that are locally faithful within the surroundings or vicinity of the observation/sample being explained.

Though LIME limits itself to supervised Machine Learning and Deep Learning models in its current state, it is one of the most popular and used XAI methods out there. With a rich open-source API, available in R and Python, LIME boasts a huge user base, with almost 8k stars and 2k forks on its GitHub repository [https://github.com/marcotcr/lime]

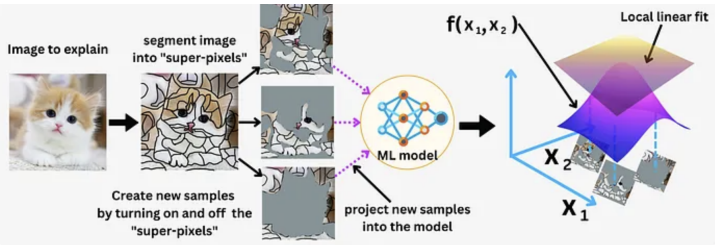

Figure 4: SmoothGradCAM

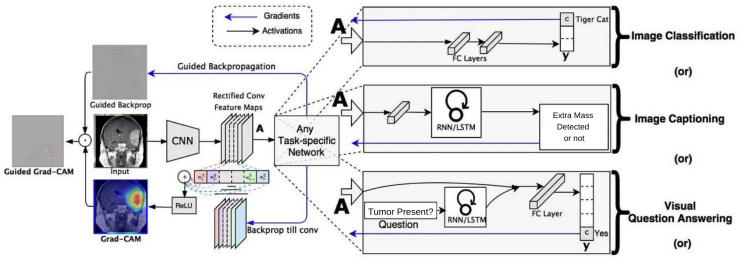

Figure 5: GradCAM flow

5. Proposed Solution: Leveraging Convolutional Neural Networks for Brain Tumor Detection

5.1. Introduction:

Introduction: Brain tumor detection and classification play a pivotal role in modern medical diagnostics and treatment planning. Timely and accurate identification of brain tumors is crucial for patient prognosis and management. With the advent of deep learning techniques, particularly Convolutional Neural Networks (CNNs), significant progress has been made in automating this process using medical imaging data.

5.2. Problem Statement:

Despite the advancements, challenges persist in achieving robust and reliable brain tumor detection systems. These challenges include handling diverse tumor types, variations in tumor

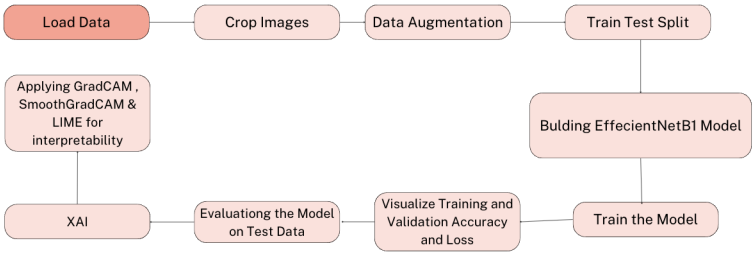

Figure 6: Methodology flow

shapes and sizes, and the need for interpretability in the model's decision-making process. Addressing these challenges is vital to develop a comprehensive solution for brain tumor detection.

5.3. Proposed Methodology:

To address the aforementioned challenges, we propose a comprehensive solution leveraging CNNs for brain tumor detection. Our proposed methodology comprises the following key components:

Data Preprocessing and Augmentation: We preprocess the brain MRI images to enhance their quality and extract relevant features. Additionally, data augmentation techniques are employed to increase the diversity and size of the training dataset, thereby improving the model's generalization capability.

Model Architecture: We adopt the EfficientNetB1 architecture, a state-of-the-art CNN model known for its efficiency and effectiveness in image classification tasks. We customize the model by adding additional layers for fine-tuning to the specific task of brain tumor detection.

Training Strategy: The model is trained using a combination of supervised learning techniques and transfer learning. We utilize a large dataset of brain MRI images, categorized into different tumor classes, for model training. During training, we employ techniques such as early stopping and learning rate scheduling to prevent overfitting and enhance convergence.

Evaluation Metrics: We evaluate the performance of the proposed model using standard evaluation metrics such as accuracy, precision, recall, and F1-score. Additionally, we analyze the confusion matrix and generate classification reports to assess the model's performance across different tumor classes.

Interpretability with GradCam: To enhance the interpretability of the model's predictions, we incorporate Gradient-weighted Class Activation Mapping (Grad-CAM) techniques. Grad-CAM generates heatmaps overlaid on input images, highlighting the regions of interest that contribute most to the model's decision-making process. This enables clinicians to understand the model's reasoning and increases trust in its predictions.

6. Results

In this section, we present the results obtained from the experimentation conducted on the EfficientNetB1 convolutional neural network (CNN) model for the classification of brain tumors. The performance evaluation encompasses training and test accuracy, confusion matrix analysis, precision, recall, F1-score, and support metrics. Additionally, we utilized explainable AI techniques including GradCAM, SmoothGradCAM, and LIME for visualizing and interpreting the model's predictions.

6.1. Model Performance Metrics

6.1.1. Training and Test Accuracy

The EfficientNetB1 model achieved a training accuracy of 99.85% and a training loss of 0.76%, while achieving a test accuracy of 95% and a test loss of 0.20%. These results demonstrate the model's ability to effectively learn and generalize patterns within the dataset.

6.1.2. Confusion Matrix Analysis

The confusion matrix provides insight into the classification performance of the model across different classes. Fig 6 illustrates the confusion matrix obtained from the test dataset:

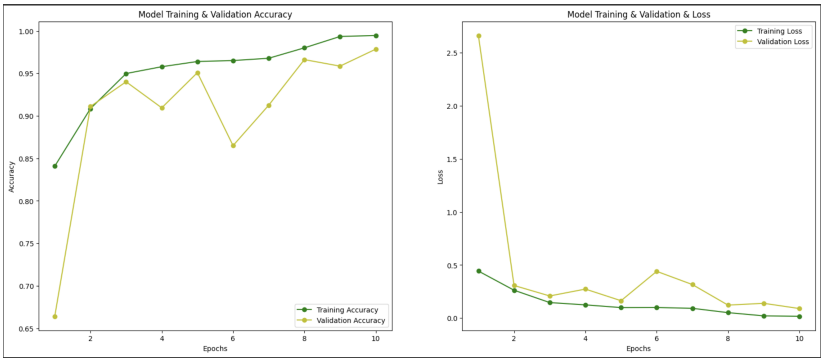

Figure 7: Model Evaluation using Accuracy and Loss

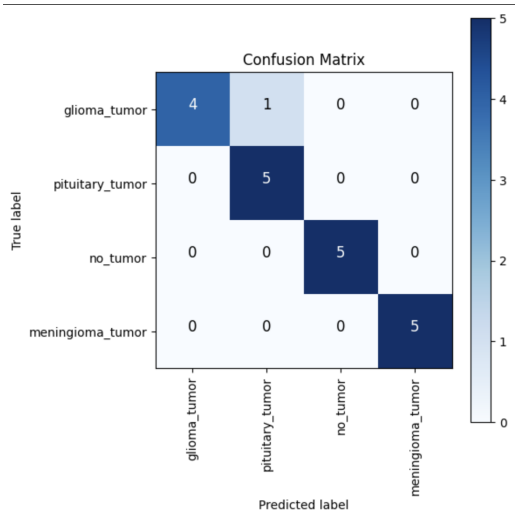

Figure 8: Confusion Matrix

6.1.3. Precision, Recall, and F1-score

Precision, recall, and F1-score are essential metrics for evaluating the model's performance on each class. Table 1 summarizes these metrics for each class:

6.2. Actual vs Predicted Class Visualization

In this section, we present screenshots illustrating the actual vs predicted class for selected samples from the test dataset. These screenshots provide visual insights into the performance of the model on individual samples.

6.3. Visual Explanations

6.3.1. GradCAM Visualization

GradCAM visualization technique was employed to highlight the regions of the input images that contributed most significantly to the model's predictions. Figure 11 and 12 presents sample

Table 1: Precision, recall, and F1-score

Class	Precision	Recall	F1-score	Support
glioma	1.00	0.80	0.89	5
meningioma	0.83	1.00	0.91	5
pituitary	1.00	1.00	1.00	5
No˙Tumor	1.00	1.00	1.00	5
Accuracy	-	-	0.95	20
Macro avg	0.96	0.95	0.95	20
Weighted avg	0.96	0.95	0.95	20

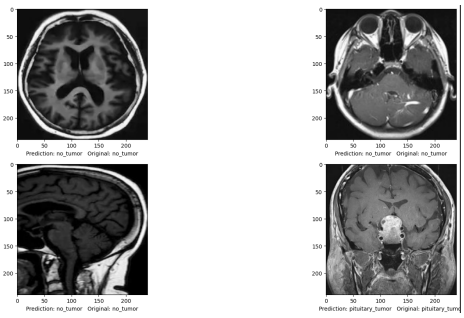

Figure 9: Enter Caption for Image 1

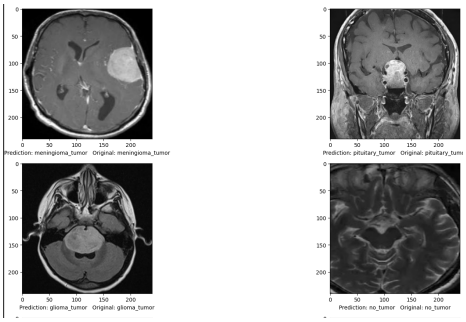

Figure 10: Enter Caption for Image 2

images overlaid with GradCAM heatmaps, providing interpretability to the model's decisions.

6.3.2. SmoothGradCAM Visualization

SmoothGradCAM was utilized to further enhance the interpretability of the model by reducing noise in the GradCAM heatmaps. Figure 13 showcases sample images with SmoothGradCAM heatmaps, offering clearer insights into the model's decision-making process.

6.3.3. LIME Explanations

Local Interpretable Model-agnostic Explanations (LIME) were generated to provide local interpretations of individual predictions. Figure 14 illustrates sample images overlaid with LIME explanations, aiding in understanding the model's behavior at a granular level.

7. Conclusion And Future Work

This study proposed an explainable Convolutional Neural Networks (CNN) model, which is using a new Grad-Class Activation Mapping (CAM) method for the brain tumor diagnosis problem. In detail, the widely used CNN technique was built with specific architecture model to ensure good enough diagnosis for dataset consisting of MRI images of brain, and the explainability level of the model was improved thanks to the explainable artificial intelligence (XAI) perspective done via Grad-CAM, SmoothGradCAM and LIME. As forming the CNN model accordingly, the developed approach was applied for the brain tumor diagnosis problem, and the model was evaluated in terms of both technical and explainability perspective.Our study

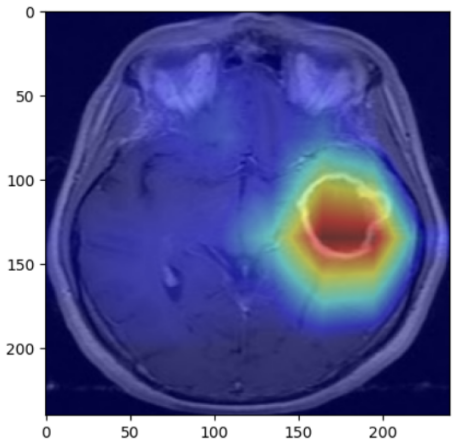

Figure 11: GradCAM Img 1

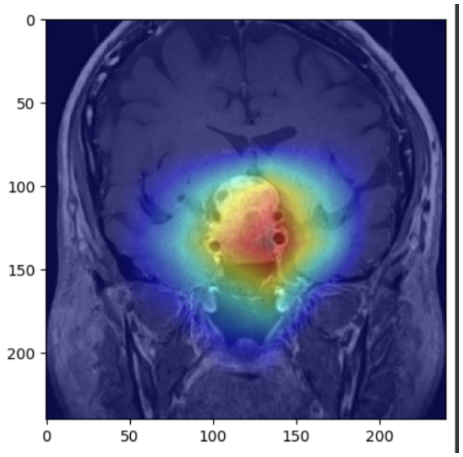

Figure 12: GradCAM Img 2

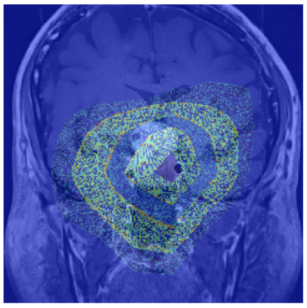

Figure 13: Interpretability using SmoothGradCAM

demonstrated that the explainable Grad-CAM method could visualize the network's performance, distinguishing the images with and without tumor based on the lesion's localization rather than other features in the brain. Thus, an enhanced visual saliency map can help increase our understanding of the internal workings of trained deep convolutional neural network models at the inference stage.

According to the obtained findings, the developed solution provided positive outcomes regarding the brain tumor diagnosis targeted in this study. Of course, there are also limitations considering open scope of brain tumor datasets and deep learning models to develop. As a result of the positive state caught in this study, the authors have already planned to go towards the mentioned limitations in future works.
Also, future works include integration of the model to different platforms (i.e. mobile platforms / devices) and different diagnosis applications (i.e. cancer, eye diseases). There are also more future works including use of other different XAI methods and compare them with the built GradCAM and other method. Also future work would involve exploring the fusion of multiple imaging modalities (such as MRI, CT, and PET scans) and non-imaging data (such as patient

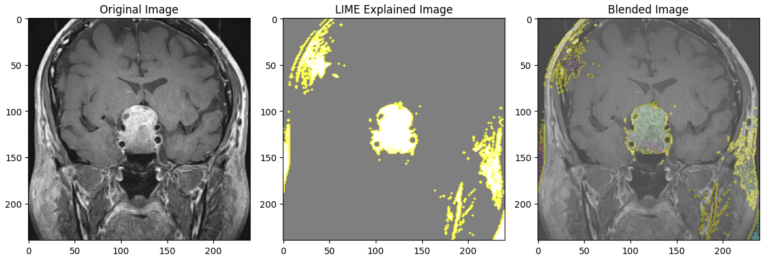

Figure 14: Interpretability using LIME

demographics and clinical history) to improve the model's robustness and generalization capabilities. Multimodal fusion techniques can enhance feature representation and capture complementary information from different data sources.
In addition to the technical, comparative evaluations for the diagnosis process, the model has to be analyzed by physicians to see if it is acceptable for decision-making support. The physicians need to attend to several tests to give feedback for explainability capabilities of the model and should decide if it is better than alternative CAM integrations. If Physicians think that the model is successful at tumor brain diagnosis and providing enough information to understand if its outcomes are as a chance or safe enough for further applications then there will be need to integrate the model to different platforms and perform similar methods for other diseases.

8. Acknowledgement

We extend our sincere gratitude to all those who have contributed to the completion of this research paper on "Explainable AI for Brain Tumor Detection". Your support, guidance, and expertise have been invaluable throughout this endeavor.
we would like to express our deep appreciation to our mentors, Dr. Abhinav Tomar, Dr. Gaurav Singal, and Dr. Vijay K. Bohat, for their unwavering support and insightful guidance.
I also want to thank my colleagues and peers for their collaborative efforts and constructive feedback, which have greatly enriched the quality of this research.
Furthermore, we are grateful to the researchers and scholars whose innovative work has laid the groundwork for our study, inspiring us to explore new possibilities in the field of explainable AI.
To everyone who has contributed in any capacity, we offer my heartfelt thanks. Your contributions have played a vital role in the success of this research endeavor.

9. References

1. Agarwal, V. (2020, May 23). Complete architectural details of all EfficientNet models. Medium. Retrieved September 19, 2021, from `https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142`.
2. Quick brain tumor facts. National Brain Tumor Society. (2021, March 22). Retrieved September 20, 2021, from `https://braintumor.org/brain-tumor-information/brain-tumor-facts/`.

3. Siddhartha. (2019, June 5). CAM visualization OF EFFICIENT. Machine Learning Blog. Retrieved September 20, 2021, from https://sidml.github.io/efficientnet-gradcam-comparison-to-other-models/.
4. Rosebrock, A. (2020, March 9). Grad-CAM: Visualize Class Activation Maps with Keras, TensorFlow, and Deep Learning. PyImageSearch. Retrieved September 10, 2021, from https://www.pyimagesearch.com/2020/03/09/grad-cam-visualize-class-activation-maps-with-keras-tensorflow-and-deep-learning/.
5. Tan, M.; Le, Q.. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97:6105–6114. Available from https://proceedings.mlr.press/v97/tan19a.html.
6. Hossain, T., Shishir, F. S., Nasim, M. A. M. A., & Shah, F. M. (2020). Brain Tumor Detection Using Convolutional Neural Network.
7. Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2020). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review.
8. Zulfiqar, Bajwa, U. I., & Mehmood, Y. (2020). Multi-class classification of brain tumor types from MR images using EfficientNets.
9. Esmaeili, M., Vettukattil, R., Banitalebi, H., Krogh, N. R., & Geitung, J. T. (2021). Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis.
10. Esmaeili, M., Vettukattil, R., Banitalebi, H., Krogh, N. R., & Geitung, J. T. (2021). Explainable Artificial Intelligence for Human-Machine Interaction in Brain Tumor Localization.