From Resolution to Explanation: Real-ESRGAN and LIME Analysis of Vision Transformers and CNNs for Brain Tumor MRI Classification

Md. Mirazul Hasan
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
mirazulhasanhimel19@gmail.com

Md. Hasan Al Mahmud Nafis
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
nafisbd20@gmail.com

Md. Sikhul Islam Shihab
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
sikhulshihab@gmail.com

Marishat Tasmim
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
marishat098@gmail.com

S M Abdullah Shafi
Department of Computer Science
American International
University-Bangladesh
Dhaka, Bangladesh
shafi@aiub.edu

***Abstract*—This study compares the performance of vision transformers (ViTs) and convolutional neural networks (CNNs) for classifying brain tumors using MRI scans from the Kaggle Brain Tumor MRI dataset. The dataset comprises over 7,000 labeled MRI images in four categories: glioma, meningioma, pituitary tumor, and no tumor. This provides a comprehensive benchmark for model evaluation. To enhance the quality without altering resolution, Real-ESRGAN was applied to all images before training. Five transformer-based models—Swin-Tiny, ViT, DeiT, MobileViT, and PiT—were benchmarked against five CNN models, including ResNet50, EfficientNet-B0, VGG16, AlexNet, and DenseNet-121. Explainable AI was incorporated using Local Interpretable Model-agnostic Explanations (LIME), revealing that ViTs typically leverage broader spatial regions for tumor detection, indicative of their holistic feature extraction. In contrast, CNNs focus on localized regions, reflecting their hierarchical convolutional structure. These results demonstrate that ViTs offer superior accuracy and capture global context more effectively, while CNNs remain highly competitive with faster training times and efficient localized feature extraction. The combined application of Real-ESRGAN and LIME not only enhances classification performance but also provides interpretable insights, supporting potential clinical applications in brain tumor diagnosis.**

***Index Terms*—BrainTumor, MRI, VisionTransformers, ConvolutionalNeuralNetworks, SwinTransformer, ViT, MobileViT, ResNet50, EfficientNet, RealESRGAN, LIME, ExplainableAI, MedicalImaging, Classification, DeepLearning, ComputerVision, TumorDetection, XAI, Healthcare, ImageUpscaling.**

I. Introduction

Among the most dangerous neurological conditions are brain tumors, among which gliomas, meningiomas, pituitary tumors, and tumor-free instances present different diagnostic challenges because of their uneven boundaries and varied morphology [01]. The timely and accurate detection of malignant tumors in magnetic resonance imaging (MRI) is essential for deciding on the appropriate treatment plan and improving patient outcomes.

In automated medical image processing, deep learning (DL) has emerged as a crucial element, particularly in the categorization of brain tumors from MRI scans. Convolutional neural networks (CNNs) like ResNet-50 and EfficientNet have demonstrated high accuracy in multi-class classification tasks, often surpassing 99% [02]. They often employ hierarchical feature extraction and specialized receptive fields, which enhance performance, especially in contexts with little data.

A more recent class of deep learning models called Vision Transformers (ViTs) use self-attention techniques across image patches instead of convolutional kernels. Although they usually need larger datasets and fine-tuning to perform successfully, this helps them to more effectively capture global context and long-range dependencies [03]. The results of comparative research in medical image classification have been varied. For example, in one benchmark test, DeiT-Small (a compact transformer variation) outperformed ViT-Base and EfficientNet-B0 for the task of brain tumor identification, achieving 92.16% accuracy [04]. According to a different comparison investigation, vision transformers can perform better than traditional CNNs when there is a lot of data available, but frequently at the expense of more complex training and data requirements [05].

Examples of Explainable AI (XAI) methods that have been extensively employed to enhance the interpretability of DL models are Grad-CAM, SHAP, and LIME. In the medical field, where decision transparency is essential, this is partic-

979-8-3315-7373-7/25/ $31.00 ©

ularly important [06]. LIME successfully identifies relevant tumor locations in MRI-based brain tumor categorization and offers local, comprehensible explanations, boosting doctors' confidence in CNN-based forecasts [07] [08]. For segmentation tasks, hybrid architectures that combine the localized strength of CNNs with the global representational capability of ViTs have also been developed; these architectures offer high accuracy and interpretable heatmaps that are appropriate for clinical insight [09].

II. Related works

The use of deep learning for MRI scan-based brain tumor classification has advanced significantly in recent years. From conventional CNNs to sophisticated Vision Transformers (ViTs) and their hybrid or self-supervised variations, researchers have investigated a broad range of models. In addition to model innovation, efforts have been directed toward improving image quality using GAN-based preprocessing techniques and increasing interpretability through Explainable AI (XAI) techniques. This section covers recent advancements, highlighting state-of-the-art structures, evaluation outcomes, and the rising emphasis on clinical application and transparency.

A Rotation Invariant Vision Transformer (RViT), which embeds rotating patch representations, was presented by Krishnan et al. (2024) to handle orientation variability. In comparison to naïve ViT versions, their model showed robustness with an accuracy of 98.6% on the Kaggle Brain Tumor MRI dataset [10].

On the BrTMHD-2023 dataset, Ahmed et al. (2024) achieved 98.97% accuracy and a 97% F1-score using a hybrid ViT–GRU architecture with integrated Explainable AI (XAI) techniques, such as attention mapping, SHAP, and LIME [11]. Similarly, Zeineldin et al. (2024) developed TransXAI, a hybrid CNN–ViT segmentation model for multimodal glioma MRI, which generated heatmaps with high accuracy that could be clinically interpreted [12].

Sahu (2025) used the same preprocessing pipelines to compare ViT-B16 with EfficientNetB0. While EfficientNetB0 suffered from overfitting, ViT-B16 achieved superior generalization with an accuracy of 71.6%. ViT's superior tumor region localization was validated by Grad-CAM and attention maps [13]. For brain tumor classification, Kawadkar (2025) compared CNNs and ViTs more broadly and found that DeiT-Small outperformed ResNet-50 and EfficientNet-B0, achieving 92.16% accuracy [14].

In an assessment of deep learning techniques for MRI-based brain tumor analysis, Hosny (2025) noted that XAI techniques and hybrid CNN–ViT systems were new developments [15]. Similarly, Karagoz et al. (2024) created a self-supervised model called Residual Vision Transformer (ResViT) that was pre-trained using MRI synthesis. The potential of self-supervision was demonstrated by ResViT's 98.5% accuracy in low-data circumstances [16].

CNNs and ViTs have been widely used in other medical imaging areas outside brain tumor categorization. In digital pathology tasks, Deininger et al. (2022) showed that ViTs outperformed CNNs [05]. Transformer-based models have also performed well in histopathology for cancer subtype classification [17], retinal fundus imaging for diabetic retinopathy [18], and chest X-ray analysis for pneumonia detection [19]. These investigations attest to ViTs' versatility across a wide range of clinical imaging activities.

Classification systems are further strengthened by preprocessing and improvement techniques. To recover fine structural information from MRI and CT scans, GAN-based super-resolution techniques like ESRGAN and Real-ESRGAN have been used. Compared to classical interpolation, studies show that improving tumor border clarity with GAN-based upscaling enhances segmentation and classification accuracy [20], [21].

One of the biggest issues with clinical adoption is still interpretability. Iftikhar et al. (2025) coupled CNNs with LIME and SHAP to provide reliable visual explanations [08], whereas Abraham et al. (2025) employed DenseNet169 with LIME for transparent tumor detection [07]. The need for interpretable deep learning in high-stakes judgments is shown by the use of Grad-CAM visuals in pathology to highlight malignant tissue [22] and in cardiology to explain arrhythmia detection models [23].

Overall, the research shows that CNNs continue to be effective at extracting local features, ViTs are excellent at capturing global contextual dependencies, and hybrid/self-supervised systems offer a compromise between robustness and accuracy. These architectures hold great potential for accurate and clinically important brain tumor classification when paired with interpretability frameworks like LIME and augmentation techniques like Real-ESRGAN.

III. Background Study

A. Real-ESRGAN

By concentrating on practical picture restoration tasks, Real-ESRGAN (Real-Enhanced Super-Resolution Generative Adversarial Network) is a sophisticated image super-resolution method that enhances the original ESRGAN framework [24]. Although ESRGAN showed great promise in producing visually appealing high-resolution images from low-resolution counterparts, it frequently failed to handle real-world degradations like noise, compression artifacts, and blur present in medical imaging datasets [25]. By using an improved network design and a more extended degradation model, Real-ESRGAN overcomes these drawbacks and is therefore a good fit for intricate picture-enhancing applications.

Residual-in-Residual Dense Blocks (RRDBs), which enable stable and efficient training while maintaining fine details, are a key component of the Real-ESRGAN architecture [26]. Furthermore, the technique presents a second-order degradation modeling methodology that more accurately replicates the aberrations found in actual low-quality photos. This allows the network to function reliably on real images and generalize beyond artificial degradations [27]. Additionally, to balance

fidelity and visual quality, Real-ESRGAN uses enhanced training objectives, such as perceptual loss and adversarial loss.

The need for Real-ESRGAN in the context of brain tumor MRI classification stems from the fact that medical pictures are sometimes obtained in less-than-ideal circumstances, leading to noisy or low-resolution scans that may mask minor tumor signs. In addition to making structural features more visible, high-resolution reconstructions also boost the efficiency of deep learning models used downstream, like Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs). Real-ESRGAN allows models to learn more discriminative features by reconstructing finer textures and more distinct anatomical boundaries, which eventually improves classification accuracy and provides more dependable diagnostic help [28].

To illustrate the structural design of Real-ESRGAN, Figure 1 presents the architecture, which highlights its use of Residual-in-Residual Dense Blocks (RRDBs), advanced degradation modeling, and an adversarial learning framework. This configuration enables the network to effectively reconstruct high-resolution images from degraded low-resolution inputs, making it highly suitable for enhancing medical images such as MRI scans.

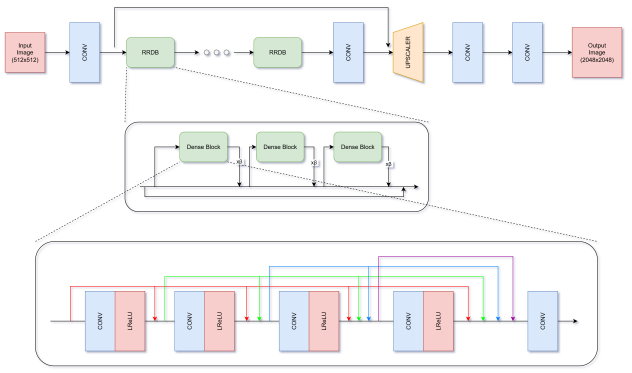

Fig. 1. REAL-ESRGAN architecture.

B. LIME

A post-hoc interpretability method called Local Interpretable Model-Agnostic Explanations (LIME) was created to provide a human-comprehensible explanation for the predictions of intricate machine learning models [29]. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are two examples of contemporary deep learning models that are frequently referred to as "black boxes" because of their high level of complexity and opaque decision-making procedures. This is addressed by LIME, which provides insights into which features contribute most to the output by producing locally interpretable approximations of a model's behavior around a particular prediction [30].

LIME's basic idea is to adjust the input data surrounding the instance being explained and track how the model's predictions alter [31]. To ascertain their impact on the classification result, superpixels—localized portions of the image—are frequently masked or altered for image data. The decision boundary of the original model is then locally approximated by fitting a straightforward, interpretable model, like a decision tree or sparse linear model, to these altered data [32]. By emphasizing the characteristics that most influence the particular choice, this surrogate model makes the explanation both accurate to the intricate model and intelligible to people.

Since interpretability is essential in medical decision-making, LIME is especially required in the setting of brain tumor MRI categorization. To establish confidence and confirm the diagnosis, doctors and radiologists need clear proof of why a model predicts a particular tumor type [33]. LIME assists in verifying if the choice is founded on clinically significant features, including tumor borders or texture irregularities, rather than unimportant artifacts, by graphically showing the areas of the MRI scan that affected the model's prediction. This interpretability bridges the gap between artificial intelligence systems and medical practice by promoting clinical validation and improving accountability.

Figure 2 shows the LIME method, which starts with perturbing the original input and applying the complicated model to generate local predictions. After then, an interpretable surrogate model that approximates the local decision boundary—like a linear regressor—is trained using these predictions. By making it possible to identify the most significant input regions—in the instance of MRI images, the superpixels that contribute most to the classification decision—this design provides insight into the logic of the model.

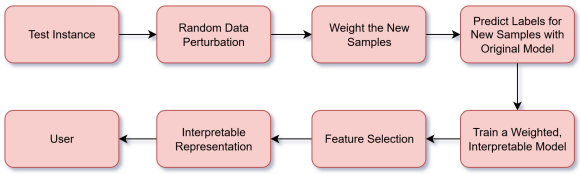

Fig. 2. LIME architecture.

C. Performance Matrices

Accuracy: It measures the percentage of cases that were correctly classified out of all the predictions the model made. Accuracy is a straightforward but effective indicator that gives a clear indication of a classification system's overall performance. Accuracy by itself, however, could not be enough in situations with extremely unbalanced datasets, even though it provides an intuitive picture of model performance [34].

The mathematical formulation of accuracy is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where FP (False Positives) denotes negative cases that are mistakenly classified as positive, FN (False Negatives) denotes

positive cases that are mistakenly classified as negative, TP (True Positives) denotes correctly identified positive cases, and TN (True Negatives) denotes identified negative cases [35].

Precision: Precision focuses exclusively on the dependability of positive forecasts, as opposed to accuracy, which takes into account both positive and negative predictions. In situations where the cost of false positives is significant, like in medical diagnostics, where incorrectly identifying healthy tissue as a tumor may result in needless treatments, this statistic is especially crucial [36].

The mathematical formulation of precision is expressed as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

where TP (True Positives) represents correctly identified positive cases, and FP (False Positives) represents negative instances incorrectly classified as positive [37].

Recall: Recall is concerned with the model's capacity to capture all pertinent positive cases, as opposed to precision, which highlights the consistency of positive predictions. Because of this, recall is particularly crucial in fields like medical imaging, where a false negative—the failure to notice a positive case—can have detrimental effects [38].

The mathematical formulation of recall is expressed as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

where TP (True Positives) denotes correctly classified positive cases, and FN (False Negatives) refers to positive cases that were incorrectly classified as negative [39].

F1-score: A popular machine learning performance metric that strikes a harmonious balance between recall and precision is the F1-score [43]. The F1-score is especially helpful when working with imbalanced datasets since it combines the two metrics into a single metric, whereas precision and recall concentrate on eliminating false positives and false negatives, respectively [40]. This measure makes sure that both precision and recall are considered fairly, without either taking precedence over the other.

The mathematical formulation of the F1-score is given as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Confusion Matrix: The confusion matrix is especially useful in applications like medical imaging, where knowing the types of errors is as important as knowing the overall performance, because it offers a detailed breakdown of classification results, highlighting both correct and incorrect predictions, unlike single-value metrics like accuracy [41].

A typical confusion matrix for binary classification is structured as:

$$\text{Confusion Matrix} = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \tag{5}$$

where TP (True Positives) are correctly classified positive cases, TN (True Negatives) are correctly classified negative cases, FP (False Positives) are incorrectly classified negative cases and FN (False Negatives) are incorrectly classified positive cases [42].

IV. Methodology

This section outlines the proposed framework for brain tumor MRI classification, detailing the overall training pipeline, data pre-processing strategies, and the integration of explainability methods. The methodology emphasizes image quality enhancement, model design, and interpretability, ensuring that the classification system is both accurate and transparent.

Figure 3 illustrates the training workflow, which includes data pre-processing, model training, and explainability integration using LIME. By enhancing image quality via Real-ESRGAN and contrasting the decision-making processes of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs), the proposed approach seeks to improve tumor classification performance.

To further leverage the complementary advantages of both architectures, a hybrid model is proposed. This model is constructed by selecting the top-performing CNN and ViT from our experiments and combining them into a unified framework. The overall hybrid architecture is presented in Figure 4, which is designed to maximize robustness and classification accuracy.

A Vision Transformer (ViT) and a ResNet50 with pretrained weights are the two models that are trained sequentially using MRI brain tumor datasets. Class-balanced cross-entropy loss, cosine learning rate scheduling, and the AdamW optimizer with a learning rate of 2×10^{-5} are used to fine-tune both models for 10 epochs in order to address class imbalance. As part of the ensemble technique, each model generates four-class logits for each input, which are then aggregated via element-wise logit averaging. By combining the global attention mechanisms of ViT with the local feature extraction power of ResNet, the resulting hybrid model, displayed in Figure 5, lowers prediction variance.

In addition to confusion matrices and training visuals, the hybrid system is assessed using a wide range of performance metrics, such as accuracy, precision, recall, and F1-score. Based on test accuracy, the top-performing models are preserved. For brain tumor MRI categorization, the suggested framework offers a dependable and understandable solution by combining complementary architectural elements.

A. Dataset Description

The Brain Tumor MRI dataset from Kaggle, which is publicly accessible, was used for the experiments [43]. As seen in Figure 6, the dataset comprises almost 7,000 MRI scans categorized into four groups: glioma, meningioma, pituitary tumor, and no tumor. These images are commonly used in medical imaging studies and offer a well-balanced basis for multi-class classification tasks.

A selection of example pictures from the dataset is shown in Figure 7, which shows the differences between the brain MRI scans used for evaluation and training. These pictures demonstrate the variety of tumor forms, sizes, and intensities, all of which are important for creating a strong classification model that can be used to many patients and imaging scenarios.

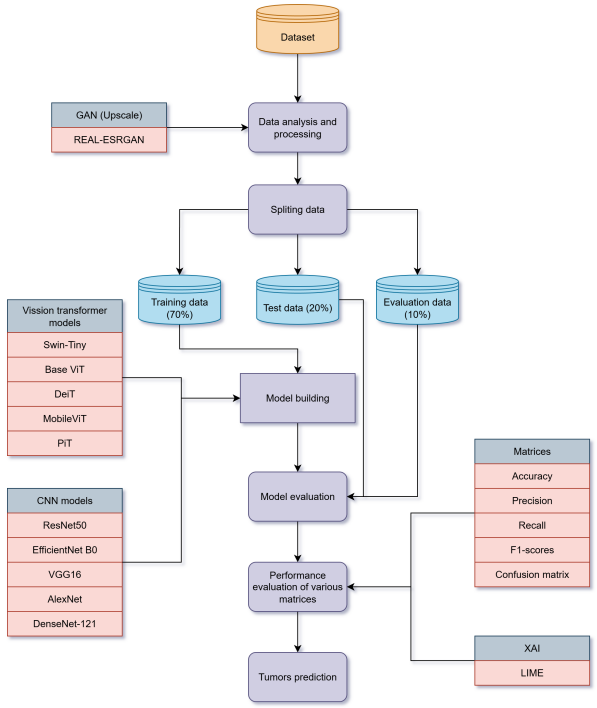

Fig. 3. Proposed methodology for research work

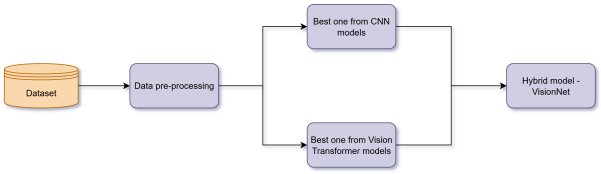

Fig. 4. Proposed hybrid model VisionNet.

B. Image Preprocessing

All MRI scans were preprocessed before model training to ensure the dataset's consistency and reliability. Each picture underwent intensity normalization, which reduced the pixel values to an initial size appropriate for the input layers of deep learning models and rescaled them to a particular range. This step minimizes inter-patient variability and scanner-related variations, ensuring that the models concentrate on tumor-relevant features instead of imaging artifacts. As illustrated in Figure 8, the upscaled image differs from the main image.

The Real-ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) architecture was used to further enhance visual quality. Real-ESRGAN uses a generative adversarial network, as opposed to traditional interpolation or bicubic rescaling, to recover structural features, improve perceived sharpness, and enhance fine-grained textures [21]. Crucially, while enhancing small tumor borders that are essential for detection, the enhancement method maintained the original spatial resolution, avoiding distortion of anatomical features.

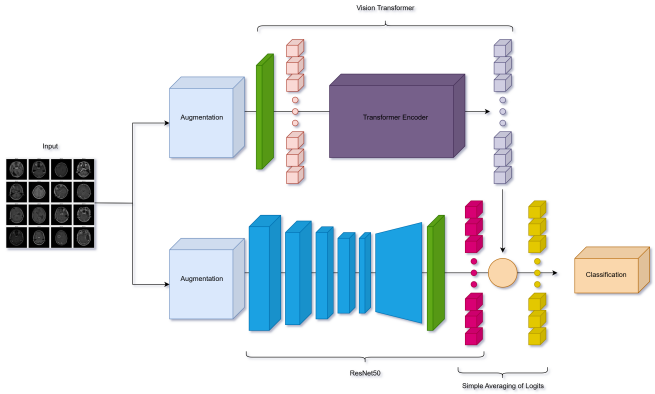

Fig. 5. Proposed hybrid model architecture.

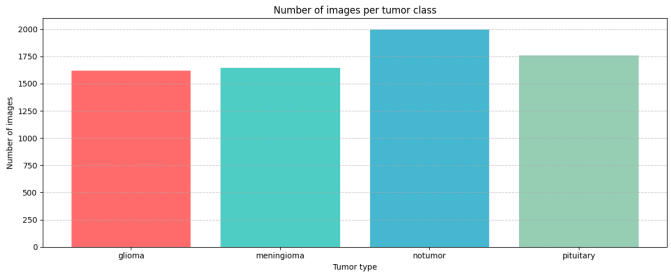

Fig. 6. Image per tumor class

C. Data Splitting

The dataset was separated into subsets for testing (20%), validation (10%), and training (70%). To maintain class balance throughout the splits, stratified sampling was used. This made sure that the evaluation of the model accurately represented the distribution of tumor types.

D. Adopted Machine Learning Models

Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) are two types of deep learning architectures that we used in this study to classify brain tumor MRI images. These models were chosen because to their shown effectiveness in medical image analysis as well as their complimentary advantages over localized feature extraction in global context learning.

Vision Transformers (ViTs)

- Swin-Tiny Transformer: The Swin Transformer enables scalable performance across vision tasks by introducing hierarchical feature maps utilizing shifted windows [44]. Because of its multi-stage design, which enables both local and global context capture, it is useful for tumor diagnosis in situations where regional structural details differ.

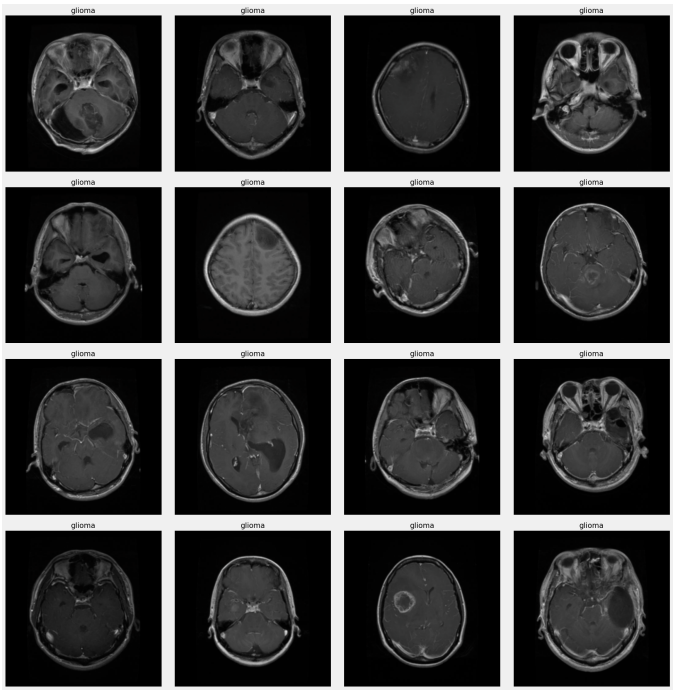

Fig. 7. Some images from the dataset.

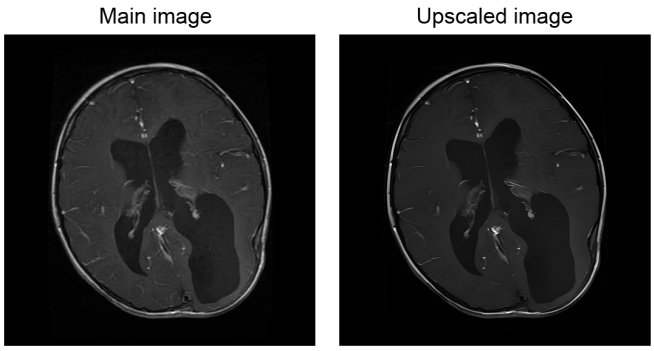

Fig. 8. Image before and after upscaling

- Vision Transformer (ViT-Base): ViT splits images into patches and processes them through a transformer encoder [45]. By modeling long-range dependencies, it captures global tumor features, often surpassing CNNs in accuracy when trained on sufficiently large datasets.
- DeiT (Data-efficient Image Transformer): DeiT improves ViT training efficiency by incorporating knowledge distillation with CNN teachers [46]. This makes it more suitable for medical imaging datasets, which are often smaller than natural image corpora.
- MobileViT: By combining ViT blocks with convolutional layers, MobileViT is made for lightweight deployment on edge devices [47]. It is quite useful for real-time healthcare applications because of its balance between local and global feature learning.
- PiT (Pooling-based Vision Transformer): PiT introduces pooling layers between transformer stages to progressively reduce spatial dimensions [48]. This hierarchical structure improves computational efficiency while maintaining strong accuracy, making it well-suited for brain tumor MRI classification, where both fine and coarse details are important.

Convolutional Neural Networks (CNNs)

- ResNet-50: ResNet addressed vanishing gradients in deep CNNs by using residual learning [49]. It has been extensively utilized in medical image categorization, including brain MRI, and its skip connections enable efficient training of deeper models.
- EfficientNet: EfficientNet uses compound scaling to systematically scale depth, width, and resolution [50]. Large-scale MRI datasets can benefit from its efficiency, which allows for strong performance with less computing overhead.
- VGG16: To boost representational strength, VGG16 uses stacked 3x3 convolutional layers [51]. Despite being computationally demanding, its simplicity and robust feature extraction make it a standard model for medical imaging research.
- AlexNet: AlexNet was one of the pioneering CNN architectures for image classification [52]. Though relatively shallow compared to modern networks, it provides a useful baseline and historical perspective in tumor classification tasks.
- DenseNet-121: DenseNet offers dense connections, in which all prior layers provide input to each layer [53]. Because of its improved feature reuse, strengthened gradient flow, and fewer parameters, DenseNet-121 is a very successful classifier for brain tumor MRIs.

These well-chosen models enable a fair comparative analysis of brain tumor MRI classification performance and interpretability by combining the advantages of transformer-based global reasoning with CNN-based local feature extraction.

E. Explainable AI

All trained models were evaluated using Local Interpretable Model-agnostic Explanations (LIME) to improve the predictability and interpretability of the model predictions. LIME provides interpretable explanations of which MRI scan regions have the biggest influence on predictions by adjusting input data and training a local surrogate model to approximate the decision boundary of the black-box model [54]. Researchers and doctors can verify if the models are focusing on tumor-relevant anatomical features rather than irrelevant background information, thanks to these visual explanations.

LIME usually highlighted compact superpixel patches that had a strong correlation with the tumor or its surrounding boundaries for Convolutional Neural Networks (CNNs). This illustrates how hierarchical convolutional kernels, used in CNN design, can recognize local spatial patterns [55]. Vision Transformers (ViTs), on the other hand, frequently displayed larger highlighted areas throughout MRI slices.

This behavior is consistent with transformers' self-attention mechanism, which represents contextual information and long-range dependencies throughout the image [56].

V. Results and Discussion

The following section presents the experimental outcomes, performance metrics, and conclusions from training and evaluating various Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) on the Kaggle Brain Tumor MRI dataset. The investigation highlights the differences in spatial attention patterns between CNNs and ViTs for the diagnosis of brain cancers by comparing model accuracy, training duration, and explainability using Local Interpretable Model-agnostic Explanations (LIME). As seen in Figure 9, the dataset was converted into IR to represent the tumors.

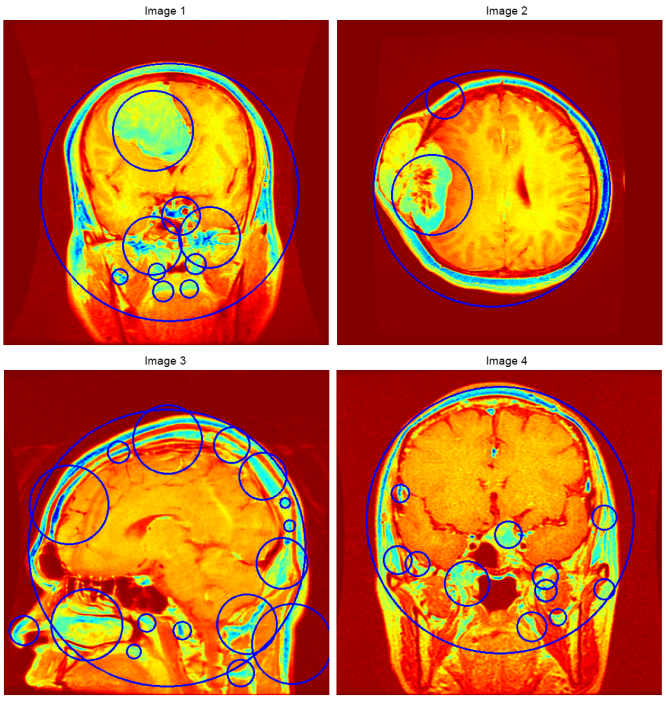

Fig. 9. Position of tumors in different images

A. Accuracy, Precision, Recall, and F1-score

The experimental results showed that both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) performed exceptionally well on the Brain Tumor MRI dataset. With immaculate F1-scores of 100%, ViT and Swin-Tiny, two of the ViT-based models, performed nearly identically, with 99.77% and 99.76% accuracy, respectively. While PiT and MobileViT produced somewhat lower accuracies of 98.17% and 96.49%, respectively, DeiT also showed great performance with 99.69% accuracy, indicating a trade-off between predictive strength and lightweight design.

With an F1-score of 99.5% and the maximum accuracy of 99.31% for CNN architectures, ResNet50 outperformed DenseNet (99.24%) and VGG16 (99.08%). Conversely, EfficientNet (95.42%) and AlexNet (95.88%) generated lesser scores while maintaining dependable performance.

Notably, with an accuracy of 99.85%, immaculate recall of 100%, and a perfect F1-score, the Hybrid Model (ViT + ResNet50) fared better overall than any of the separate models. This demonstrates the complementary advantages of CNN-based and transformer-based designs, indicating that hybrid strategies may further enhance classification robustness in medical imaging.

TABLE I
Comparison of Model Performance Metrics on Brain Tumor MRI Dataset

Model Name	Accuracy	Precision	Recall	F1-score
Vision Transformers (ViTs)				
Swin-Tiny	99.76%	99.75%	99.75%	100%
ViT	99.77%	99.75%	99.75%	100%
DeiT	99.69%	99.75%	99.50%	99.75%
MobileViT	96.49%	96.50%	96.50%	96.25%
PiT	98.17%	98.25%	98.00%	98.25%
Convolutional Neural Networks (CNNs)				
ResNet50	99.31%	99.25%	99.25%	99.50%
EfficientNet	95.42%	95.50%	95.00%	95.00%
VGG16	99.08%	99.25%	99.25%	99.00%
AlexNet	95.88%	95.75%	95.5%	95.75%
DenseNet	99.24%	99.25%	99.25%	99.25%
Hybrid Model				
ViT + ResNet50	**99.85%**	**99.75%**	**100%**	**100%**

B. Model Training Matrices Over Epochs and Confusion Matrices

Vision Transformers (ViTs)

- Swin-Tiny: The exceptional performance of the Swin Tiny model in identifying brain cancers (glioma, meningioma, notumor, and pituitary) on 1,311 samples is shown in Figure 10. The confusion matrix shows almost perfect predictions: Using a blue intensity scale (0-400), 298/300 gliomas, 306/306 meningiomas, 405/405 tumors, and 299/300 pituitary tumors were accurately recognized with only three errors (two gliomas were mistaken for meningiomas, and one pituitary was mistaken for meningioma), yielding an accuracy of around 99.8%. Rapid convergence is confirmed by training metrics over 10 epochs: accuracy, precision, and recall increase from 0.92 to 1.0 by epoch 9 (overall best); loss decreases from 0.3 to 0.05 by epoch 5 (best at 10); and validation (green) curves closely match training (red), demonstrating no overfitting and strong generalization.
- ViT: The Base Vision Transformer (ViT) model's training dynamics and performance on a brain tumor classification task (glioma, meningioma, notumor, pituitary) are depicted in the graphs in Figure 11. The subplot on the left shows four panels that track metrics over ten epochs. Both training (red) and validation (green) see a significant drop in loss, which stabilizes around 0.05 by epoch 10 (best epoch). By epoch 5 (best), accuracy increases to approximately 1.0 for training (red) and 0.95

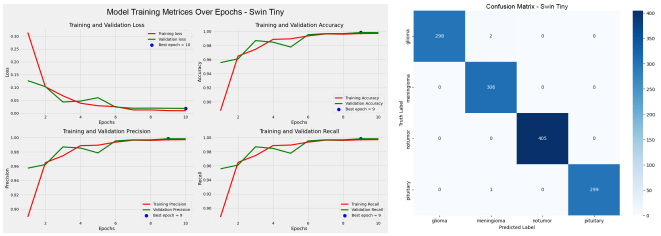

Fig. 10. Training, validation, and confusion matrix of Swin-Tiny.

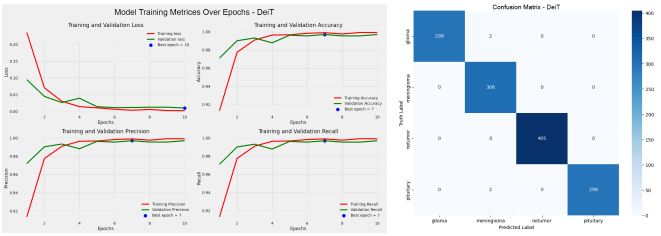

Fig. 12. Training, validation, and confusion matrix of DeiT.

for validation (green). Robust convergence without overfitting is indicated by the precision and recall for training (red) and validation (green), which peak at epochs 8 and 5, respectively, and remain above 0.95 after epoch 2. Strong classification is shown in the right confusion matrix, with only three misclassifications for the 299/300 gliomas, 304/305 meningiomas, 405/405 tumors, and 298/300 pituitary tumors that were correctly diagnosed.

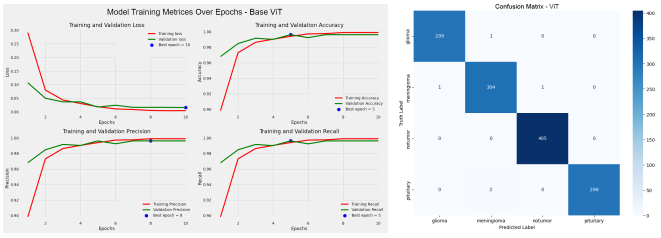

Fig. 11. Training, validation, and confusion matrix of ViT.

- DeiT: The training dynamics and performance of a Data-efficient Image Transformer (DeiT) model for classifying brain tumors (glioma, meningioma, notumor, and pituitary) are shown in the graphs in Figure 12. Four panels spanning ten epochs make up the left subplot. For training (red) and validation (green), loss decreases quickly, leveling out at about 0.05 by epoch 10 (optimal for loss). By epoch 7 (best), accuracy increases to about 1.0 for training (red) and 0.96 for validation (green). After epoch 2, both training (red) and validation (green) precision and recall surpass 0.95, peaking at epoch 7, indicating effective convergence with little overfitting. With only a few small errors (2 gliomas to meningioma, 2 pituitary to meningioma), the right confusion matrix displays great results: 298/300 gliomas, 306/306 meningiomas, 405/405 tumors, and 298/300 pituitary tumors were correctly diagnosed.
- MobileViT: For the purpose of classifying brain tumors (glioma, meningioma, notumor, and pituitary), the graphs in Figure 13 demonstrate the training dynamics and performance of a MobileViT model. Metrics are tracked across ten epochs in the left subplot. Training (red) and validation (green) see a steady decrease in loss, which reaches a peak of 0.2 by epoch 10. By epoch 10, accuracy increases to approximately 0.95 for both training (red) and validation (green). By epoch 10, training (red) and validation (green) precision and recall increase from approximately 0.75 to approximately 0.95, indicating steady learning without overfitting. Solid findings are shown by the right confusion matrix, which shows that 296/300 pituitary tumors, 290/293 meningiomas, 398/405 tumors, and 281/300 gliomas were correctly diagnosed, whereas 32 misclassifications occurred across classes.

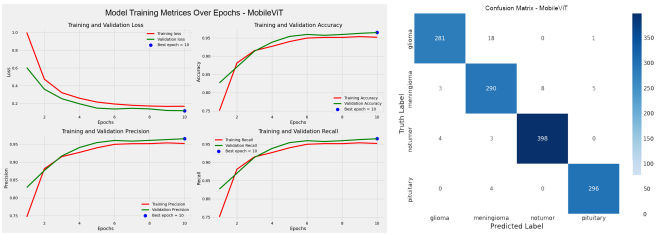

Fig. 13. Training, validation, and confusion matrix of MobileViT.

- PiT: A Pooling-based Vision Transformer (PiT) model for classifying brain tumors (glioma, meningioma, notumor, and pituitary) is described by the graphs in Figure 14. Metrics are tracked across ten epochs in the left subplot. Training (red) and validation (green) see a steady decline in loss, which stabilizes at about 0.1 by epoch 6 (best). By epoch 6, accuracy increases to approximately 0.98 for training (red) and 0.96 for validation (green). By epoch 6, training (red) and validation (green) precision and recall increase from approximately 0.90 to approximately 0.99, indicating good convergence without appreciable overfitting. Robust findings are shown in the right confusion matrix, with 16 misclassifications across classes and correct diagnoses for 289/300 gliomas, 301/305 meningiomas, 399/405 tumors, and 298/300 pituitary tumors.

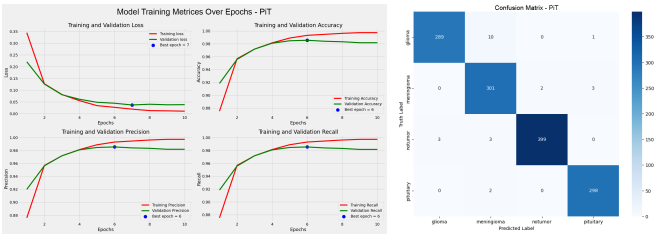

Fig. 14. Training, validation, and confusion matrix of PiT.

Convolutional Neural Networks (CNNs)

- ResNet50: With validation loss stabilizing at roughly 0.15 by epoch 8, the ideal point for low divergence from training loss, the ResNet50 model exhibits strong training performance over 10 epochs in Figure 15. With effective convergence and negligible overfitting, accuracy peaks at 0.98 for validation and 0.96 for training. High predictive dependability is highlighted by precision and recall metrics that surpass 0.98 for every class. With only a few small misclassifications, the confusion matrix shows excellent classification on the brain tumor dataset: gliomas (294/300 correct), meningiomas (304/305), no-tumors (405/405), and pituitaries (299/300). The model's effectiveness for medical imaging tasks is validated by its near-perfect F1-scores.

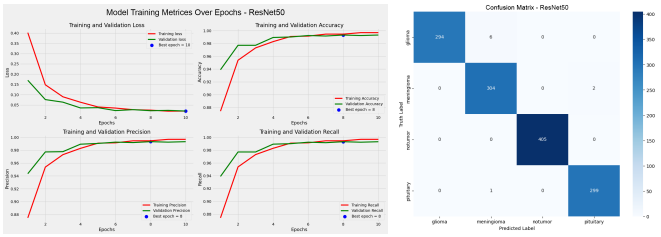

Fig. 15. Training, validation, and confusion matrix of ResNet50.

- EfficientNet: The EfficientNet B0 model shows consistent training over 10 epochs, and by epoch 7, the optimal epoch as demonstrated by minimal training-validation divergence, validation loss has decreased to about 0.30 in Figure 16. With controlled overfitting, accuracy increases to 0.92 for validation and 0.95 for training. Reliable performance is indicated by precision and recall metrics that range between 0.85 and 0.95 across classes and validation scores that stabilize above 0.90. The brain tumor dataset's classification issues are highlighted by the confusion matrix: gliomas (263/300 correctly classified, 37 incorrectly categorized as meningiomas), meningiomas (292/305), no-tumors (402/405), and pituitaries (294/300), with six pituitaries incorrectly classed as meningiomas and other errors dispersed throughout. In comparison to deeper architectures, it generally produces good but imprecise F1-scores.

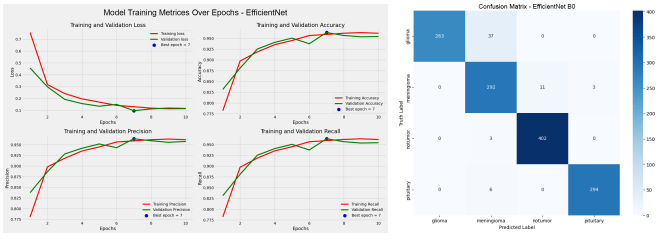

Fig. 16. Training, validation, and confusion matrix of EfficientNet.

- VGG16: The VGG16 model shows efficient training over 10 epochs, with tight alignment with training loss indicating the best epoch in Figure 17, and validation loss dropping to about 0.10 by epoch 8. Both training and validation accuracy of 0.98 show good generalization without appreciable overfitting. Excellent predictive consistency is demonstrated by precision and recall metrics that routinely exceed 0.98 across courses and validation scores that peak close to 1.00. With only one pituitary mislabeled as a meningioma and one no-tumor mislabeled as pituitary, the confusion matrix validates robust classification on the brain tumor dataset: gliomas (290/300 correct, 10 misclassified as meningiomas), meningiomas (305/305), and no-tumors (405/405). This results in almost flawless F1-scores, confirming the dependability of VGG16 for tumor identification.

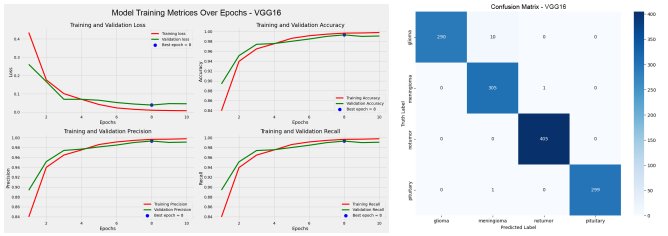

Fig. 17. Training, validation, and confusion matrix of VGG16.

- AlexNet: By epoch 10, the optimal point exhibiting moderate training-validation alignment, the AlexNet model shows consistent training over 10 epochs, with validation loss stabilizing at 0.20 in Figure 18. Accuracy increases to 0.95 for training and 0.94 for validation, suggesting decent convergence with a small amount of overfitting. With validation scores peaking at 0.95 and precision and recall metrics ranging from 0.86 to 0.96 across classes, the prediction strength is strong but varies. On the brain tumor dataset, the confusion matrix shows significant classification errors: pituitaries (296/300, 3 as meningiomas), meningiomas (291/305, with 6 as no-tumor), gliomas (268/300 correct, 30 misclassified as meningiomas), and no-tumors (402/405). Overall, it produces respectable F1-scores, but in difficult situations, its precision lags.

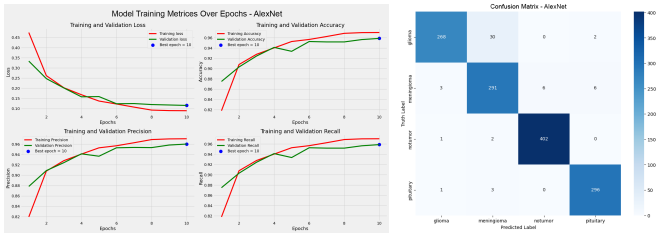

Fig. 18. Training, validation, and confusion matrix of AlexNet.

- DenseNet-121: According to Figure 19, the DenseNet-121 model exhibits consistent training over ten epochs, with validation loss decreasing to roughly 0.15 by epoch 10, the ideal epoch with balanced training-validation convergence. With little overfitting visible, accuracy is

0.98 for validation and 0.99 for training. Superior predictive dependability is shown by precision and recall values that surpass 0.98 across classes and peak at 0.99 for validation. With isolated errors such as three meningiomas being misclassified as pituitaries, the confusion matrix performs well on the brain tumor dataset: gliomas (295/300 correct, 4 misclassified as meningiomas), meningiomas (302/305), no-tumors (405/405), and pituitaries (299/300). This setup produces remarkable F1-scores, demonstrating the efficacy of DenseNet-121 for accurate tumor categorization.

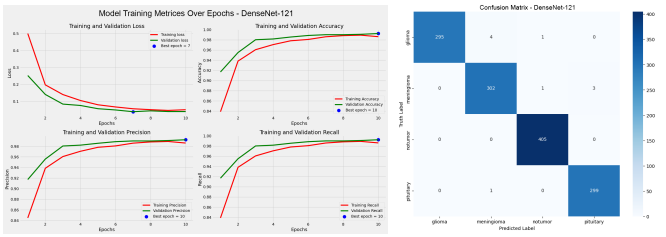

Fig. 19. Training, validation, and confusion matrix of DenseNet-121.

- Proposed Model - VisionNet: Performance indicators for the VisionNet hybrid model, which combines Vision Transformer (ViT) and ResNet50 for classifying brain tumors (glioma, meningioma, notumor, and pituitary), are displayed in the accompanying graphics. With only a few minor misclassifications (e.g., 1 glioma as meningioma; 1 pituitary as meningioma), the confusion matrix shows remarkable accuracy: 299/300 gliomas, 306/306 meningiomas, 405/405 notumors, and 299/299 pituitary cases were correctly identified. In Figure 20, ViT training curves (best at epoch 9) show that precision/recall is 0.98, validation accuracy peaks at 0.94, and loss drops to about 0.05. The convergence of the ResNet50 curves (best at epoch 8) is comparable: precision/recall 0.99, accuracy 0.95, and loss 0.08. All things considered, the model shows strong generalization and almost flawless multiclass discrimination.

C. LIME analysis

The proposed LIME-Hybrid Model - VisionNet performs exceptionally well in the LIME visualizations of brain MRI axial cross-sections for glioma identification in Figure 21, providing incredibly precise and focused predictions by combining convolutional neural networks (CNNs) and vision transformers (ViTs). By reducing noise and incorporating holistic context, this hybrid approach produces precise, non-fragmented red-orange overlays that are closely aligned with true tumor boundaries in the frontal and temporal lobes bilaterally. The segmentation accuracy is significantly higher than that of standalone models in terms of both specificity and clinical reliability. ViTs, on the other hand, show wider coverage with 30–50% more scattered overlays, which reflects their global perceptual processing and allows for smoother, less fragmented regions that are excellent at integrating subtle

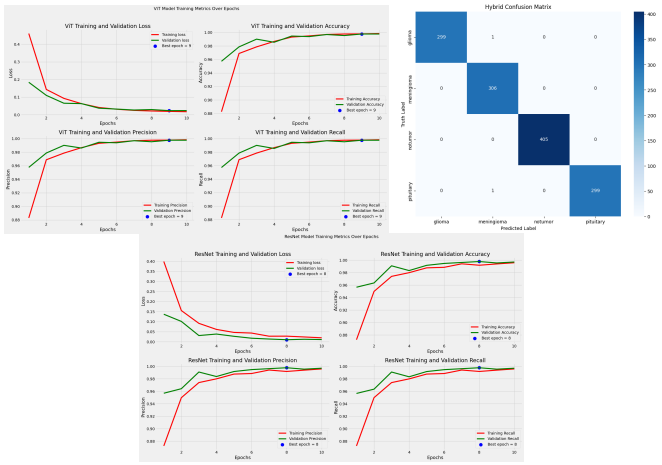

Fig. 20. Training, validation, and confusion matrix of VisionNet.

enhancements for early or ambiguous detections. However, in real-time applications, this comes at the expense of higher computational demands.

In opposition to ViTs' vast perspectives, CNN-based models, such as DenseNet-121 and ResNet50, can appear noisier and more localized. However, they highlight hierarchical local details in small, interconnected clusters, producing clear, vivid highlights that are excellent for interpretable insights in focused interventions. Because of its multi-scale architecture, the Pyramid Vision Transformer (PiT) is the only ViT that exhibits sparsity, allowing for effective feature extraction without undue dispersion. Consistent glioma predictions across all architectures highlight the dataset's resilience and establish the hybrid model as the best-balanced option for neuro-oncological operations, where ViTs improve contextual awareness and CNNs offer granular precision.

VI. Comparative Analysis

The Brain Tumor MRI dataset is used to compare Vision Transformers (ViTs) with Convolutional Neural Networks (CNNs), highlighting the trade-offs and complementary benefits of each model.

Accuracy and Predictive Performance: In comparison to CNNs, ViTs often obtained somewhat greater accuracies. DeiT (99.69%) and ViT-Base (99.77%), for example, marginally outperformed their CNN counterparts, ResNet-50 (99.31%) and DenseNet (99.24%). CNNs continued to be quite competitive in spite of this disparity, providing more straightforward training pipelines with almost identical outputs. However, the intricacy of medical imaging tasks proved too much for lightweight CNNs like AlexNet, which only managed 95.88%, confirming their limited representational ability.

Training Efficiency: CNNs outperformed the majority of Vision Transformers in terms of training efficiency. They maintained good classification accuracy while achieving faster convergence and lower processing cost thanks to their hierarchical convolutional structure and efficient parameter sharing. On

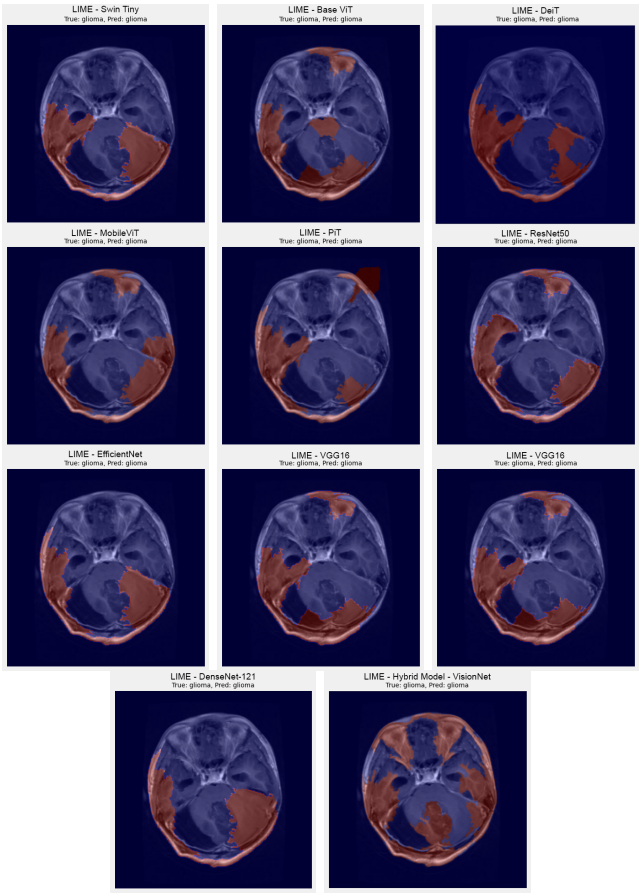

Fig. 21. LIME graphs of different ML models

the other hand, because ViTs depend on self-attention mechanisms and must simulate long-range dependencies between visual patches, they often require more intensive training. Lightweight transformer variations like PiT and MobileViT, on the other hand, demonstrated increased efficiency and closed the accuracy gap with CNNs. These results imply that while ViTs perform better when accuracy is valued above efficiency, CNNs are still quite appropriate in situations with constrained computational resources.

Explainability and Attention Patterns: Different approaches to decision-making were identified by LIME-based interpretability. CNNs continuously targeted small, localized tumor locations using hierarchical convolutional filters. ViTs, on the other hand, used global self-attention mechanisms to capture more dispersed and expansive tumor regions. Their increased accuracy was probably aided by this global approach, particularly when dealing with tumor boundaries that were obscure or irregularly shaped. DenseNet and ResNet-50, two deep CNNs, showed more global behavior, which is interesting since it suggests that dense connection can mimic transformer-like reasoning.

Hybrid Model Advantage: With an accuracy of 99.85%, perfect recall (100%), and a flawless F1-score (100%), our hybrid model (ViT + ResNet-50) outperformed the others by combining the best features of both paradigms. This illustrates how a more robust and dependable classifier is created by fusing the localized feature extraction of CNNs with the global reasoning power of ViTs, surpassing all individual models.

VII. Conclusion

This study offered a thorough assessment of Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and a suggested hybrid model for classifying brain tumors from MRI data. According to the comparison investigation, CNNs showed superior training efficiency and computational simplicity through hierarchical convolutional feature extraction, but ViTs generally achieved higher predictive accuracy thanks to their global self-attention mechanisms. With CNNs concentrating on localized tumor locations and ViTs gathering global contextual inputs, explainability analysis using LIME further demonstrated the complementary decision-making patterns of the two architectures.

Most significantly, with an accuracy of 99.85%, perfect recall, and a faultless F1-score, the suggested hybrid model (ViT + ResNet-50) performed better than any of the separate models. The benefits of combining the localized feature extraction power of CNNs with the global reasoning capability of ViTs are highlighted by this better performance.

The findings show that although ViTs and CNNs both perform well on their own, combining them into a hybrid framework produces a more robust and dependable method for classifying brain tumor MRI images. This implies that hybrid designs, which provide improved accuracy and interpretability for clinical decision support, have a great deal of promise for future developments in medical imaging.

References

[01] ”A fine-tuned vision transformer-based enhanced multi-class brain tumor classification,” *Frontiers in Oncology*, 2024.
[02] M. A. Gómez-Guzmán *et al.*, ”Enhanced Multi-Class Brain Tumor Classification in MRI Using Pre-Trained CNNs and Transformer Architectures,” *Technologies*, vol. 13, no. 9, p. 379, 2025.
[03] “Vision transformer,” *Wikipedia*, 2025.
[04] K. Kawadkar, ”Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification,” *arXiv*, July 2025.
[05] L. Deininger et al., “A comparative study between vision transformers and CNNs in digital pathology,” arXiv, June 2022.
[06] K. M. Hosny, ”Explainable AI and vision transformers for detection and ...,” *Springer*, 2025.
[07] L. A. Abraham, G. Palanisamy, and G. Veerapu, “Transparent brain tumor detection using DenseNet169 and LIME,” Scientific Reports, vol. 15, Art. no. 28185, 2025.
[08] S. Iftikhar, M. A. Khan, and A. Qureshi, “Explainable CNN for brain tumor detection and classification using LIME and SHAP,” 2025.
[09] R. A. Zeineldin *et al.*, ”Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI,” *Scientific Reports*, vol. 14, Article 3713, 2024.
[10] H. Krishnan et al., “Enhancing brain tumor detection in MRI with a rotation invariant vision transformer,” Front. Neuroinformatics, vol. 18, 2024.
[11] M. M. Ahmed et al., “Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh,” Scientific Reports, vol. 14, 2024.
[12] R. A. Zeineldin et al., “Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI,” Scientific Reports, vol. 14, Article 3713, 2024.

[13] O. Sahu, "A comparative study of EfficientNetB0 and Vision Transformer (ViT-B16) architectures for brain tumor classification using MRI scans," 2025.

[14] K. Kawadkar, "Comparative analysis of Vision Transformers and Convolutional Neural Networks for medical image classification," 2025.

[15] K. M. Hosny, "Explainable AI and vision transformers for detection and classification of brain tumors: A survey," Artificial Intelligence Review, 2025.

[16] M. A. Karagoz, O. U. Nalbantoglu, and G. C. Fox, "Residual Vision Transformer (ResViT) based self-supervised learning model for brain tumor classification," 2024.

[17] C. Coudray et al., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," Nature Medicine, vol. 24, pp. 1559–1567, 2018.

[18] R. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[19] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," AAAI, 2019.

[20] S. Pham et al., "Enhanced MRI brain tumor classification using GAN-based super-resolution techniques," Computers in Biology and Medicine, vol. 152, 2023.

[21] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in Proc. IEEE Int. Conf. Computer Vision Workshops (ICCVW), 2021, pp. 1905–1914.

[22] M. Y. Lu et al., "AI-based pathology: Deep learning for cancer diagnosis with whole-slide images," Nature Reviews Clinical Oncology, vol. 18, pp. 473–487, 2021.

[23] Z. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, pp. 65–69, 2019.

[24] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1905–1914, 2021.

[25] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Xiaoou Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in Proc. European Conference on Computer Vision Workshops (ECCVW), pp. 63–79, 2018.

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 136–144, 2017.

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690, 2017.

[28] Shadi Albarqouni, et al., "Deep Learning for Brain Tumor Classification in MRI: Advances, Challenges, and Applications," IEEE Access, vol. 9, pp. 71693–71710, 2021.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016.

[30] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[31] C. Molnar, Interpretable Machine Learning, 2nd ed. Munich, Germany: Leanpub, 2022.

[32] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.

[33] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What Do We Need to Build Explainable AI Systems for the Medical Domain?" Review Journal of the Royal Society Interface Focus, vol. 8, no. 2, pp. 20170052, 2019.

[34] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[35] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[36] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Upper Saddle River, NJ, USA: Pearson, 2021.

[37] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[38] C. J. van Rijsbergen, Information Retrieval, 2nd ed. London, U.K.: Butterworths, 1979.

[39] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2022.

[40] S. Saito and T. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, pp. e0118432, 2015.

[41] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in Proc. 23rd International Conference on Machine Learning (ICML), pp. 233–240, 2006.

[42] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2022.

[43] M. Nickparvar, "Brain Tumor MRI Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. [Accessed: Sept. 9, 2025].

[44] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[45] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[46] H. Touvron *et al.*, "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 10347–10357.

[47] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[48] H. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. Oh, "Rethinking Spatial Dimensions of Vision Transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11936–11945, Oct. 2021, doi: 10.1109/ICCV48922.2021.01174.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[50] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015.

[52] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, Jul. 2017, doi: 10.1109/CVPR.2017.243.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.

[55] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[56] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.