

Multi-Class Brain Tumor Classification from MRI Images Using Pre-Trained CNNs: Feature Representation and Uncertainty Analysis

Mousa Alalhareth 

Department of Information Systems, College of Computer Science and Information Systems,
Najran University, Saudi Arabia

Abstract—Brain tumor classification based on Magnetic Resonance Imaging (MRI) is a challenging task due to the fact that there are many kinds of brain tumors that have quite similar visual features, and there are many different variations in individual cases of the same kind of brain tumor. The aim of this work is to investigate the performance of ImageNet pre-trained convolutional neural networks (CNNs) to learn and classify deep feature representations for multi-class brain tumour classification, in addition to analysing the uncertainty of the prediction. The EfficientNetB0, ResNet50 and DenseNet121 models were trained on a total of 7,023 MRI images from four classes – Glioma, Healthy, Meningioma and Pituitary. Feature-norm analysis, principal component analysis (PCA), and class-centroid distances were used to investigate the learned feature space. The accuracy, precision, recall, F1-score, and Matthews correlation coefficient (MCC) were used to evaluate the model performance. Moreover, Monte Carlo (MC) Dropout was applied to investigate prediction confidence, prediction entropy and variance. The extracted feature representation had 1280 dimensions, with 371, 577 and 963 principal components representing 90%, 95% and 99% variance, respectively. The between-to-within class separability ratio was still relatively low, 0.148, suggesting that the learned feature space for the classes had a high intra-class variance. ResNet50 model gives the best results for classification with an accuracy of 96.11%, precision of 95.90%, recall of 95.89%, F1 score of 95.89%, and MCC of 0.94, among the three. The uncertainty analysis further demonstrated that the predictive entropy tends to increase when the model produces incorrect predictions. For ResNet50, the model achieved a mean prediction confidence of 0.91, with predictive variance of 0.006, suggesting relatively high confidence and low variability in its predictions. The results demonstrate that classification accuracy alone is insufficient to fully characterize the predictive performance of deep learning model behavior. The analysis of the learned feature space with traditional classification metrics and uncertainty measures provides a more reliable informative characterization of CNN-based brain tumor classification from MRI data.

Keywords—Brain tumor; deep learning models; feature extraction; magnetic resonance imaging

I. INTRODUCTION

Brain tumors are complex diseases of the brain and are often difficult to accurately diagnose due to the variety of their appearance and characteristics from patient to patient. Brain images can be difficult to interpret because of the differences in

size, shape, location and internal structure of the tumours [1][2]. Magnetic Resonance Imaging (MRI) is a frequently used technique in the evaluation of brain tumors because it can give in-depth information about the soft tissue of the brain and allow detection of abnormal tissue within the brain. Although there are these benefits, there are several cases where it is hard to interpret MRI scans accurately without a lot of experience on the matter. Sometimes tumors may look like normal tissues, and other tumors can look alike on MRI. These difficulties have intensified the demand for computer-aided diagnostic techniques that can help clinicians to automatically detect and classify patterns in brain MRI [3].

Artificial Intelligence (AI) is becoming more prevalent in the medical field, creating a new opportunity for analyzing medical images. Machine Learning (ML) and Deep Learning (DL) methods have gained significant interest in particular due to their capability to recognize complex patterns that are hard to describe with classical image processing techniques. Manual features, such as texture, shape, intensity or edge information, are commonly used in traditional image classification methods [4]. The usefulness of these features may significantly rely upon the quality of the feature engineering process and the image characteristics of the images to be analyzed. Deep learning, specifically Convolutional Neural Networks (CNNs), offers another solution that learns to extract valuable image representations directly from training data. CNN-based models can learn progressively about low-level features like edges and textures, and then aggregate them into more high-level, meaningful features. As a result, deep learning is increasingly important in various tasks of medical imaging of the brain, such as tumor detection, segmentation, and classification [5][6].

Previous research has explored a wide range of deep learning strategies for automated brain tumor analysis. They include classic CNN designs, transfer learning, attention-based networks, ensemble networks, recurrent networks, generative adversarial networks and more recently Vision Transformer based networks. The growing number of these techniques demonstrates the continued quest for better automated system performance to discriminate between tumor classes and to obtain accurate results from complex MRI data. But, accurate classification is still difficult, especially if there are visual similarities between classes of tumors or if there is a lot of variation within the same class [7].

While deep learning has proven to be a promising method for brain tumor analysis, there are several key challenges that remain to be addressed. A primary concern is that the patterns and representations learned by various pretrained CNN designs can vary from MRI data. It can be difficult to compare models across different studies due to differences in datasets, preprocessing techniques, training conditions and evaluation procedures. This means that different architectures should be studied under similar experimental conditions to make a fair comparison. The other drawback is that most indicators, including accuracy, F1-score, etc. are reported to assess how good a model is at prediction. These measures do not necessarily reflect the reliability of the confidence placed in such predictions. Furthermore, minimal attention has been paid to analyzing the feature representation of deep CNN models. A learned feature space can provide valuable insights into the separability of different tumor categories and reveal whether the extracted feature representations exhibit meaningful distinction between classes. These restrictions underscore the importance of having a systematic approach to assessing the quality of classification, the confidence in prediction, the uncertainty of predictions, the calibration of classifiers, and the nature of features learned under a single experimental paradigm [8][9][10].

This study aims to develop and systematically evaluate a transfer-learning-based framework for multiclass brain tumor classification using MRI images. Three popular CNN models (EfficientNetB0, ResNet50 and DenseNet121) are evaluated under the same settings to quantify their classification performance fairly. Model selection is done using the validation Macro-F1 score, while a held-out test set is used for final performance evaluation. In addition to conventional classification metrics, Monte Carlo (MC) Dropout is applied to examine predictive uncertainty through repeated multiple stochastic predictions. These predictions are used to measure confidence, predictive entropy, and variance. The predicted probabilities are further evaluated using Expected Calibration Error (ECE), Brier Score and Negative Log-Likelihood (NLL). The study also examines deep features extracted using pretrained EfficientNetB0 to explore their distribution across the four classes. Together, these analyses provide a broader evaluation of the model's performance, predictive uncertainty, probability calibration, and the characteristics of extracted feature representations. To achieve these objectives, this research is guided by the following two research questions:

RQ1: Is the degree of representational learning of discriminative features better, worse or comparable in CNN models pre-trained on ImageNet for multi-class brain tumor classification?

RQ-2: Can MC Dropout-based uncertainty estimation be used to give meaningful insight into low confidence or ambiguous brain MRI predictions?

The key findings of the study can be summarized as follows:

- A systematic comparison of transfer-learning models is implemented which is used to compare the performance of three models: EfficientNetB0, ResNet50, and DenseNet121 for the multiclass brain tumor classification problem using MRI images. An identical

experimental procedure enables the models to be compared with each other in identical conditions.

- The study of learned feature representations investigates learned feature representations extracted by EfficientNetB0 to gain insight into the structure of the learned representation space. The extent of the separation of the different tumor categories in terms of Principal Component Analysis (PCA), class-centroid distances, within-class dispersion and between-class-to-within-class separability is explored.
- The performance of the models is evaluated by multiple complementary metrics such as accuracy, precision, recall, F1 score, and (MCC). Different metrics can be combined to get a more comprehensive view of the classification performance when class-wise performance varies.
- The reliability of probabilistic predictions of the selected model and the model's calibration are analyzed with the help of the following metrics: Expected Calibration Error (ECE), Brier Score, and Negative Log-Likelihood (NLL). Reliability diagrams are also used to assess agreement between the model's predicted confidence level and the observed frequency of correct predictions.

The remainder of the paper is divided into four sections. In Section 2, the key research gap was identified. In Section 3, the overall methodology adopted was discussed. In Section 4, the interpretation of experimental analysis was defined and discussed. Finally, in Section 5, necessary future directions and research conclusions were discussed.

II. LITERATURE REVIEW

Brain tumour analysis has been a widely studied field of deep learning, for classification, segmentation, and even uncertainty estimation. Segmentation is used to locate the tumor and define the boundaries of the tumor, but is not the same as image-level classification. Therefore, the main focus of the current study is on the past studies regarding brain tumor classification, transfer learning, predictive uncertainty and probability calibration.

Accurate diagnosis and treatment planning are crucial for brain tumor. To facilitate tumor segmentation and predictive uncertainty estimation, a deep learning framework based on Residual U-Net (Res-UNet) with (MC) Dropout was developed. The residual connections were employed to allow feature learning in the deeper layers, and MC Dropout was used to estimate epistemic uncertainty. This uncertainty mapping produced uncertainty maps that showed areas of lesser reliability in the model that can be helpful for clinical interpretation of the segmentation results. The proposed method resulted in the Tversky index of 91.68% on the Kaggle Brain MRI dataset of the LGG [11]. Uncertainty quantification (UQ) has attracted increasing research interest in recent years, as high segmentation or classification accuracy does not necessarily guarantee reliable prediction for individual cases. This paper [12] analyzed MC Dropout for the diagnosis of glioma in a multi-task setting where the segmentation was combined with the tumor subtyping. The effects of the number of MC samples and the chosen drop-out

rate were specifically investigated and epistemic and aleatoric uncertainty were treated separately. Results showed that MC Dropout can yield uncertainty estimates, which are well calibrated if the model parameters are chosen properly. The reliability of uncertainty estimates in automatic brain tumor segmentation has also been investigated using several widely adopted uncertainty quantification methods. These techniques have been subjected to one study which assessed their usefulness for calibration, for providing localization of errors in segmentation, and for detecting failed segmentations. Methods provided good overall calibration at the dataset level, but at the individual-subject level some methods were found to be miscalibrated, and the error localization at the voxel level was relatively weak. Despite this, voxel-wise uncertainty measures were able to convey some value for the identification of subject-level segmentation issues [13]. Some other difficulties in automated segmentation are tumor heterogeneity and uncertainty in tumor boundary. In this study [14], a hybrid PA-ResNet50 framework was introduced for explainable panoptic brain tumor segmentation. This method was able to produce accurate tumor boundaries through a semantic and instance segmentation process and provide interpretability. The framework showed high accuracy of 99.3% and 98.7% while providing Dice scores of 99.29% and 98.85% on the BraTS 2019 and BraTS 2021 datasets, respectively. It is especially challenging to accurately segment the tumor sub-regions in cases of irregular and poorly defined boundaries. To solve this problem, an explicit incorporation of boundary information into the segmentation process was proposed; to incorporate the multi-modal fusion network, the Boundary-aware and Uncertainty-driven Multi-modal Fusion Network (BUFNet) was proposed. The framework also exploited the complementary information across various MRI sequences via multi-modal fusion. Tumour-region discrimination improvement and uncertainty estimation were also added to provide information of confidence that can help in clinical interpretation [15].

in addition to individual segmentation frameworks, there have been several studies that have explored at the broader use of AI in brain tumor diagnosis. A single study looked at (ML) and (DL) methods for automatic diagnosis and highlighted their impressive capacity to learn complex patterns in images directly from MRI data. The review also included a discussion on new developments in diagnostic algorithms and areas that should be explored in the future [16]. A more comprehensive survey was undertaken on artificial intelligence techniques for detection of brain tumors in various medical imaging modalities. It encompassed several methods, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), hybrid networks, vision transformers, transfer learning, attention models, and ensemble models. The review also compiled the data sets and methods that are used to preprocess the data in this area, as well as the evaluation measures that have been applied to various studies. Some issues that plague AI-based brain tumor analysis were also pointed out [17]. There have also been some hybrid segmentation and classification strategies investigated. In another work, fuzzy thresholding was fused with a deep neural network, where fuzzy segmentation was first employed to segment healthy and abnormal tissue and an optimized CNN was then used to classify the tumour. The four categories

included were glioma, meningioma, pituitary tumor, and no tumor. Three publicly available datasets with over 23000 MRI images were used for evaluation, leading to 98% classification accuracy and a Dice similarity coefficient of 97.97% [18]. Computational efficiency is also an important factor to consider, especially when such a model is to be used in practice. A framework of a PDSCNN-RRELM-based light network was proposed for the classification of four classes: brain tumor, glioma, meningioma, no tumor, and pituitary tumor. To enhance the visibility of the tumor, the image was preprocessed using contrast limited adaptive histogram equalization (CLAHE), and then the discriminative feature was extracted by PDSCNN with relatively low computational cost. The precision, recall and accuracy values of the RRELM-based classifier were 99.35 %, 99.30%, and 99.22%, respectively on average. SHAP was also added to give more depth to the model's prediction [19]. Another popular method of brain tumor classification is the transfer learning approach. In this study [20], the performance of some pre-trained CNN architectures like Xception, MobileNetV2, InceptionV3, ResNet50, VGG16 and DenseNet121 was compared on the publicly available Brain Tumor MRI dataset, which consists of 7023 images. Xception achieved the best performance amongst the tested architectures, with a weighted accuracy of 98.73% and a weighted F1 score of 95.29%. The outcomes once again show the ability of CNNs trained in advance to classify MRI images of glioma, meningioma, pituitary tumors, and no-tumor.

III. METHODOLOGY

This paper introduces a deep learning algorithm for analyzing MRI scans of brain tumors. The primary aim is to investigate whether pre-trained CNN model can learn meaningful image representations and discriminate among the various tumor categories. The four main stages of the proposed experimental procedure are described in Fig. 1. The first is the organization and preparation of the MRI images. The second stage consists of image preprocessing and controlled augmentation to enhance the generalization of the models while preserving key diagnostic elements from the scans. In the third stage, the learned CNN features are analyzed in detail and examined using feature extraction, dimensionality analysis, principal component analysis (PCA), class-centroid analysis, and within-class and between-class separability measurements. At the end, three of the pretrained models for transfer learning (EfficientNetB0, ResNet50, DenseNet121) are trained and tested. Classification performance is evaluated using a range of complementary metrics to provide a more comprehensive and reliable comparison of the investigated architectures.

A. Data Collection

This study used brain tumor MRI images downloaded from a publicly available source [21]. Fig. 2 represents some images from the dataset. The dataset was comprised of four classes (Glioma, Healthy, Meningioma, and Pituitary). In total, there were 7,023 image samples. The per-class sample distribution is illustrated in Fig. 3. The healthy class has more samples (2000) as compared to the others. The fewer samples (1621) were from Glioma class. Class weights were also considered in training to reduce the impact of variations in the number of samples within classes.

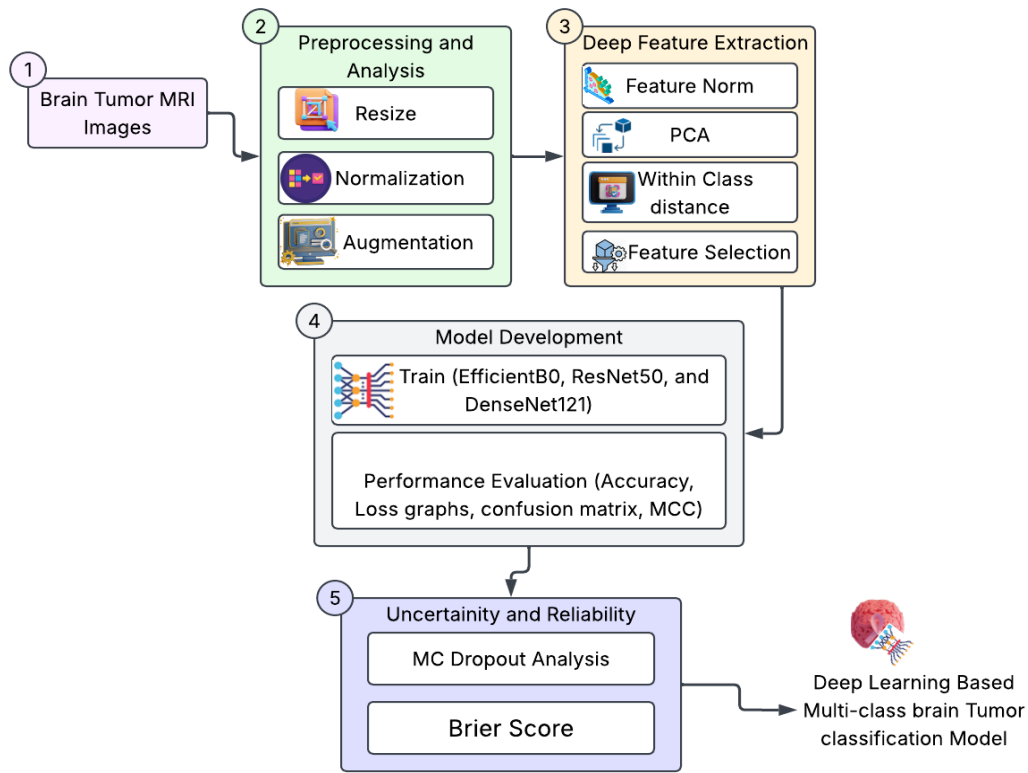


Fig. 1. Overall research pipeline.

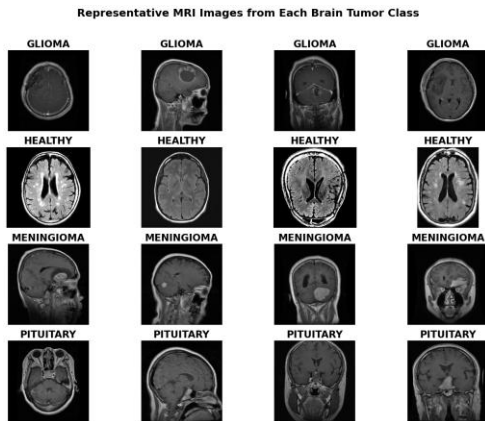


Fig. 2. Some samples from dataset.

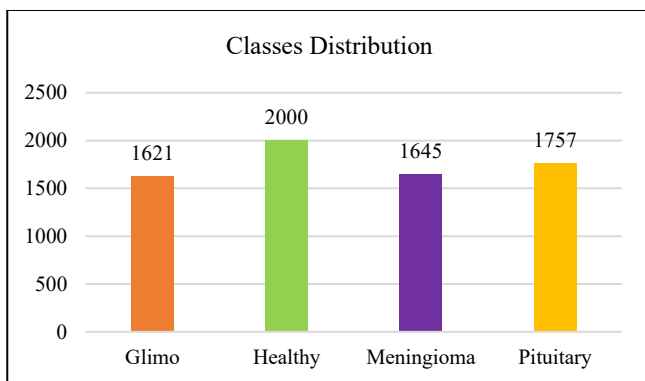


Fig. 3. Class distribution.

B. Preprocessing and Analysis

All MRI scans were resized to a fixed size of 224×224 pixels prior to feeding the models. This image size was selected to be compatible with the input images of the pretrained CNN architectures, but not so large that it would be prohibitively expensive to compute it. The MRI images were converted to RGB images because the networks selected were pretrained on ImageNet and must have three channels as input. This made it possible to have a uniform input structure for all images in the different models utilized in the experiments [22].

Data augmentation was applied to all training images to increase the models' performance in generalization to images not seen during training on unseen MRIs. Small geometric alterations such as up to 15° rotations, horizontal and vertical shifts of up to 10% and zooming up to 10% were introduced in the original images during the augmentation process. Flipping horizontally was also performed [23]. These transformations filled in empty regions around the image borders, which were then completed with nearest neighbor filling. The augmented version of an image I is:

$$I' = T(I; \theta) \quad (1)$$

where, T is the augmentation and the parameters θ are the random variables that are specified with the augmentation.

The validation and test sets were not changed during the experiments. This ensures that the reported results are based on the true classification accuracy of the trained models when they are tested on the original MRI images, and not on a modified version of the images.

C. Deep Feature Extraction

The learned representations of the MRI images were then analyzed using an ImageNet pre-trained EfficientNetB0 network prior to conducting the classification experiments. To use the network as a feature extractor instead of as an ImageNet Classifier, the original classification layer was removed. The output of the last convolutional layer can be represented by feature maps. Let the final conventional representation be expressed as:

$$\mathbf{F} \in \mathbb{R}^{\mathbb{H}_f \times \mathbb{W}_f \times \mathbb{C}_f} \quad (2)$$

where, $\mathbb{R}^{\mathbb{H}_f \times \mathbb{W}_f \times \mathbb{C}_f}$ denotes the height, width, and number of feature channel respectively. These feature maps were then averaged across the entire map, known as global average pooling (GAP), to form a **fixed**-length vector. The pooled value is defined as the value for the k -th feature channel.

$$\mathbf{z}_k = \frac{1}{\mathbb{H}_f \mathbb{W}_f} \sum_{i=1}^{\mathbb{H}_f} \sum_{j=1}^{\mathbb{W}_f} \mathbf{f}_{i,j,k} \quad (3)$$

where, $\mathbf{z}_k$ is the value of the k -th feature in the representation.

As a result, all MRI images were transformed to a small deep feature vector. The weights of the EfficientNetB0 backbone were pre-trained on ImageNet and remained frozen during this analysis. The feature vectors were extracted from the training, validation and test sets to allow analysis of the features properties of each set.

1) *Deep feature distribution analysis:* Feature representations were then examined to see if images from various tumor classes were located in different areas of the feature space after the feature extraction. Three complementary analyses were considered: feature magnitude, feature standardization and dimensionality reduction.

2) *Feature norm analysis:* The Euclidean norm was used to measure the value of each extracted feature vector:

$$\|\mathbf{z}_i\|_2 = \sqrt{\sum_{j=1}^d \mathbf{z}_{ij}^2} \quad (4)$$

Here, d is the number of the feature dimensions.

For the four tumor classes, these resulting norm values were tested individually. This analysis gives an idea of whether the magnitude of the learned representations would differ between the various categories.

The overall distribution of feature norms for the four classes is quite similar as illustrated in Fig. 4, indicating that the representations extracted have similar values, though the different classes may have different values.

3) *Feature standardization:* Standardization was done for the extracted features prior to PCA. For each feature dimension, the standardized value was obtained as

$$\mathbf{z}'_{ij} = \frac{\mathbf{z}_{ij} - \mu_j}{\sigma_j} \quad (5)$$

where, μ_j and σ_j are the mean and the standard deviation computed over dimension j of the features, respectively.

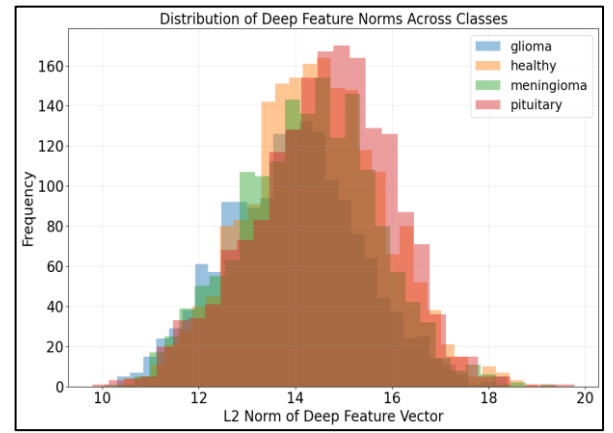


Fig. 4. Deep feature norm across various classes.

Standardization prevents the feature dimensions with large numerical scales from having a disproportionate influence from the larger dimensions in the PCA or distance calculations.

4) *Principal component analysis:* The dimensional structure of the extracted features was analyzed using Principal Component Analysis (PCA), and how the various classes of tumors were spread in a lower-dimensional feature space was assessed [24]. Assume that the standardized feature matrix is denoted by Z with Principal matrix W . The resulting representation is given by:

$$P = ZW \quad (6)$$

However, the proportion of the total variance accounted for by the X^{th} component is expressed by:

$$M_X = \frac{\lambda_M}{\sum_{j=1}^d \lambda_j} \quad (7)$$

where, λ_M is the eigenvalue corresponding to the X^{th} principal component.

The cumulative explained variance was then calculated as:

$$CEV_X = \sum_{M=1}^n M_X \quad (8)$$

The number of components required to obtain 90%, 95% and 99% of the total variance was noted. Furthermore, the first two principal components were combined to form a "two-dimensional view" of the feature space. This was projected and the clustering of samples within each class visually investigated to see if each class was represented by a distinct or overlapping cluster of samples.

Fig. 5 represents the cumulative explained variance, which shows that the obtained PCA components capture a significant amount of the information in the original feature space with a relatively small number of components.

Fig. 6 shows the two-dimensional PCA projection, which provides a graphical representation of the samples in the learned feature space. The overlap and separation between the classes observed provide a hint about how discriminable the classes are on the feature level.

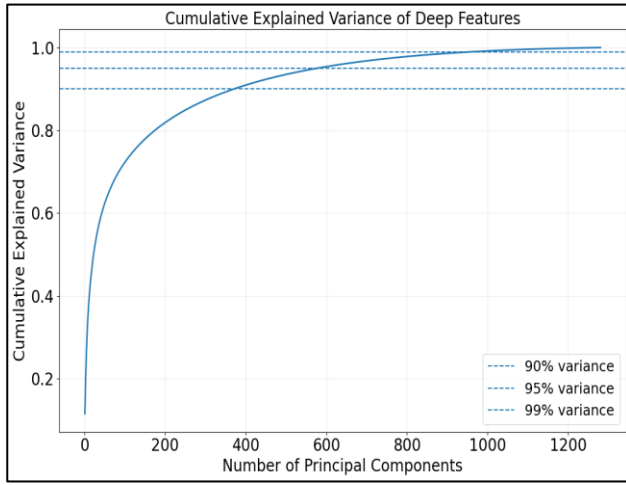


Fig. 5. Cumulative explained variance.

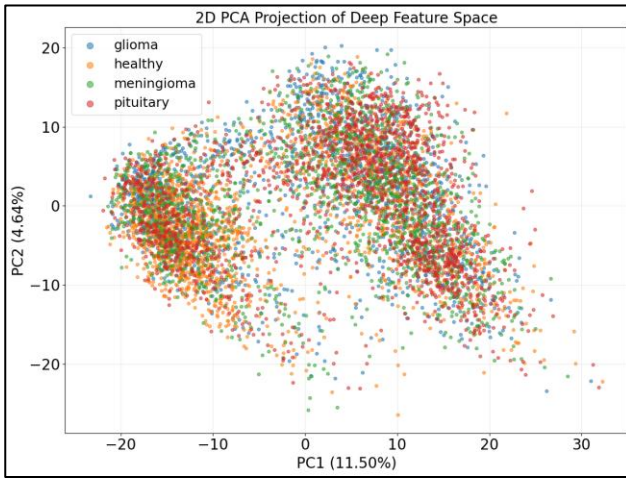


Fig. 6. PCA analysis.

5) *Class-centroid analysis*: Class centroids were used to get more insights into the representation of the position of those four tumor classes in the learned feature space. The centroid of a class $\mathcal{C}$ is calculated as

$$\mu_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{i: y_i = \mathcal{C}} \mathbf{z}_i \quad (9)$$

Here $N_{\mathcal{C}}$ represents the number of samples in class $\mathcal{C}$.

A distance between two classes $\mathcal{C}$ and $\mathcal{A}$ is computed based on the Euclidean distance:

$$\mathbb{D}(\mathcal{C}, \mathcal{A}) = \|\mu_{\mathcal{C}} - \mu_{\mathcal{A}}\|_2 \quad (10)$$

All pairs of classes were compared for distance, and the results were put into a centroid-distance matrix. The further apart the two centroids are, the further apart their respective learned representations are, while the closer together the two centroids are, the closer together their corresponding classes are, or the more similar the learned representations of the two classes are.

Fig. 7 illustrates the relative separation between the class representations, as indicated by the pairwise centroid distances.

The further apart the centroids the further apart the classes are believed to be in the learned feature space.

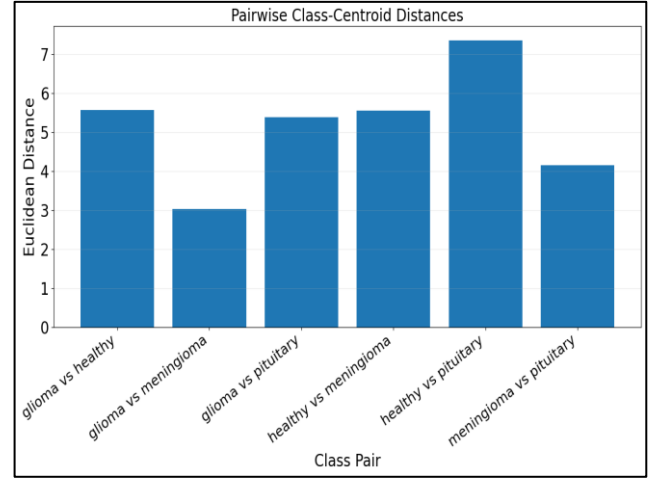


Fig. 7. Pairwise class-centroids.

TABLE I. FEATURE ANALYSIS SUMMARY

Metrics	Values
Total Samples	7,023
Feature dimension	1280
Mean feature value	0.1095
Feature Std Deviation	0.387
Mean feature Norm	14.32
90% PCA Components	371.00
95% PCA Components	577.00
99% PCA Components	963.00
Mean within-class distance	34.912
Mean between class distance	5.174
Between/within separability ratio	0.148

Table I, an overall statistical overview of the feature space of 1280 dimensions generated by the deep network, is provided. The analysis is performed based on 7,023 samples, each sampled feature vector being a 1280-dimensional vector. The mean feature value from the entire feature set is 0.1095 with a standard deviation of 0.387. The feature norm for the classes is rather stable and equals 14.32 on average. The PCA results give more information about the dimensionality of these representations. Among the 1,280 features, the 371, 577 and 963 features resulted in retaining 90%, 95%, and 99% of the variance, respectively. This implies that a lot of the information in the original feature space is contained in fewer principal components, following which the dimensions of the rest of the feature space add up comparatively less to the variance in the data. The distance-based statistics show a different property of the feature space. The mean within-class distance is 34.912, which is significantly larger than the mean between class distance of 5.174. As a result, the between-to-within class separability ratio is just 0.148. In practical terms, samples from the same class may be widely scattered in the sample space, but centroids of the classes may be fairly grouped together. This

means the extracted representations are not well clustered and well-separated into class clusters. Thus, additional classification and feature-refinement steps are crucial for training decision boundaries which are capable of handling the significant intra-class variation.

D. Model Development

For the experiments, three well-known pretrained CNN architectures were chosen: EfficientNetB0, ResNet50, and DenseNet121. The weights of each network were initialized with weights from the ImageNet dataset, while the ImageNet classification layer was removed before training.

All the convolutional layers of the three networks were not updated, while just the newly added classification head was trained to classify the four classes of brain tumors. By keeping the backbone fixed, the study could benefit from what has been learned from the visual features pre-trained on ImageNet without tuning millions of parameters.

This configuration also allowed for the experimental conditions to remain the same for each of the three architectures, and limited the amount of computing power and the number of parameters that needed to be optimized.

1) *EfficientNetB0*: To achieve a good trade-off between the complexity of the model and computational efficiency, EfficientNetB0 was chosen as a baseline transfer-learning model. In this experiment, the classification layer from the original ImageNet network was deleted from the pretrained network [25]. The backbone's output of convolutional feature maps was fed to the classification head employed in this work. The resulting architecture is described in Fig. 8.

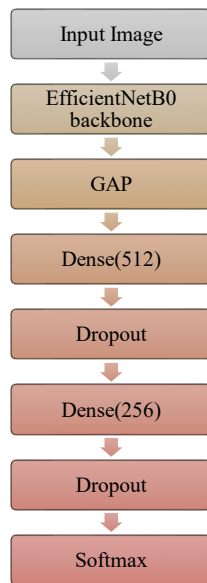


Fig. 8. Architecture of the EfficientNetB0 model,

2) *ResNet50*: The ResNet50 model is used as a baseline to investigate the performance of Residual Learning in Brain Tumor Classification. The convolutional layers were pretrained and were used as feature extractors, with the original ImageNet output layer discarded and another classification layer tailored

for the four classes of interest. The main feature of ResNet50 is that it has an architecture of residual connections, which facilitates the passage of information and gradients through deeper layers of the network. This facilitates the learning of useful features of images while not making the optimization process unnecessarily hard. For consistency with the different backbone networks, the classification head employed for ResNet50 is the same as for the other baseline models. This makes it possible to interpret any discrepancy in classification results directly as a discrepancy in the representations of the features learned by the different backbone architectures [26].

3) *DenseNet121*: To show the contribution of dense connectivity to the classification task, another baseline model, DenseNet121, was used. Unlike CNNs, DenseNet makes all connections with the subsequent layers, which means that the features derived in one layer can be used in the following layers. This connectivity is also helpful to keep the gradients stable while training. The convolutional backbone DenseNet121, which was already trained on ImageNet, was kept, and the image classification layer trained on ImageNet was discarded for the experimental task. The training was performed on the backbone alone, and the classification head was attached, but not frozen, in a later stage [27].

EfficientNetB0, ResNet50 and DenseNet121 were trained using the same experimental setup. A fair ground for comparison of feature representations generated by the three pretrained CNN backbones is the use of a common classifier and training configuration.

E. Monte Carlo Dropout Uncertainty Estimation

Deeper neural networks can give high confidence to a prediction when the input data is ambiguous, hard to classify, or close to the decision boundary. Therefore, accuracy alone does not give a full picture of the reliability of the predictions of a model. In this research, predictive uncertainty is measured in terms of Monte Carlo (MC) Dropout, which views dropout as a form of stochastic inference instead of toggling it off during training. For inference, the same input is fed into the network several times using a different dropout mask each time. This yields a set of slightly different predictions for the same sample. The range of such predictions can then be used to quantify the uncertainty of the model, such as by confidence, predictive entropy and predictive variance. These measures give a more informative evaluation of the reliability of the prediction than does accuracy alone. For an input sample x_i , the stochastic prediction at the t^{th} forward pass is defined as:

$$\mathcal{P}^{(t)}_i = f_{\theta_{m^{(t)}}}(x_i); \quad t = 1, 2, \dots, T \quad (11)$$

where, m^t represents a randomly sampled dropout mask and $T = 20$ denotes the number of stochastic forward passes. The MC predictive probability is obtained by averaging the stochastic predictions:

$$\bar{P}_i = \frac{1}{T} \sum_{t=1}^T \mathcal{P}^{(t)}_i \quad (12)$$

The final class prediction is given by:

$$\hat{Y}_i = \arg \max_K \bar{P}_i k \quad (13)$$

The prediction confidence is measured by:

$$C_i = \max_k \bar{P}_i k \quad (14)$$

While the predictive entropy is calculated as:

$$H_i = -\sum_{k=1}^K \bar{P}_i k \log(\bar{P}_i k + \epsilon) \quad (15)$$

where, ϵ prevents the numerical instability. Higher entropy indicates greater uncertainty in predicted class distribution. The variability among stochastic predictions is further quantified using predictive variance:

$$v_{ik} = \frac{1}{T} \sum_{t=1}^T (\mathcal{P}^{(t)}_{ik} - \bar{P}_i k)^2 \quad (16)$$

The classification head is made of 512 units dense layer, batch normalization, and dropout, followed by a 256 units dense layer and dropout. The dropout rates for both dropout layers are 0.30. For (MC) Dropout, 20 stochastic forward passes per test batch were made for each model with 'training=True'. The final class was determined using the average of the predicted probabilities across each of the passes and predictive entropy and variance were used to measure uncertainty. EfficientNetB0, ResNet50, and DenseNet121 were used to ensure consistency across the three models and the same procedure was applied.

One drawback is that the number of forward passes was set at 20 and no attempt was made to evaluate if a higher number would result in more stable uncertainty estimates. Therefore, this setting should not be regarded as a threshold for convergence. Additionally, including the argument 'training=True' can have an impact on the behaviour of batch-normalization and/or dropout, and thus on the predictions and uncertainty estimates. The impact of stochastic passes should be investigated in future research, and it is important to confirm that dropout is not turned off during inference, but does not cause the behaviour of batch normalization to change.

F. Experimental Setup

All experiments were conducted using python language and Kaggle environment. Table II shows the various training parameters of each model performed during training. The dataset was divided into 70% training, 15% testing and 15% validation using stratified sampling at the image level. Patient level identifiers were not available in the data set, however, and it was not possible to ensure that images from the same patient were stored in separate subsets.

TABLE II. CONFIGURATION PARAMETERS OF MODELS

Configuration	EfficientNetB0	ResNet50	DenseNet121
Input Size	224 x 224 x3	224 x 224 x3	224 x 224 x3
Pretrained Weights	ImageNet	ImageNet	ImageNet
BackBone Parameters	$\cong 5.29 M$	$\cong 23.59 M$	$\cong 7.04 M$
Feature Extraction	Conv backbone	Residual Blocks	Dense Blocks
Global pooling	GAP	GAP	GAP
Dropout	0.30	0.30	0.30
Output classes	4	4	4

Activation	Softmax	Softmax	Softmax
Optimizer	Adam	Adam	Adam
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch Size	16	16	16
Maximum Epochs	50	50	50
Loss Function	Cross Loss Entropy	Cross Entropy Loss	Cross Loss Entropy
Training samples	4916	4916	4916
Testing Samples	1054	1054	1054
Validation samples	1053	1053	1053

IV. RESULTS AND DISCUSSION

The performance of the three baseline deep learning models is evaluated using widely adopted classification metrics. Each model is classified into the different classes and the following are used to measure its performance: Accuracy, precision, Recall, F1-score, Matthew's correlation coefficient (MCC).

A. Performance Evaluation of Baseline Model

Table III shows the performance evolution of baseline models across various metrics. The ResNet50 model achieved highest overall performance accuracy (96.11%), precision (95.90%), recall (95.89%), F1 score (95.89%), and MCC score of 0.94 followed by DenseNet121 achieves an accuracy of 95.26%, precision of 95.10%, recall of 95.09%, F1 score of 95.06%, and MCC score of 0.93. EfficientNetB0 model attained an accuracy of 93.55%, precision of 93.51%, recall of 93.21%, F1 score of 93.26%, and MCC score of 0.91. ResNet50 outperforms as compared to other models with a higher F1 score and MCC, indicating it may have a better tradeoff between precision and recall, and be able to predict the classes more consistently. From these results, it is evident that ResNet50 is the best of the three baseline models considered in this study, but cannot be much better than the other architectures, since no statistical testing has been conducted to validate the differences observed.

TABLE III. PERFORMANCE EVALUATION OF MODELS

Models	Accuracy	Precision	Recall	F1 score	MCC
EfficientNetB0	93.55%	93.51%	93.21%	93.26%	0.91
ResNet50	96.11%	95.90%	95.89%	95.89%	0.94
DenseNet121	95.26%	95.10%	95.09%	95.06%	0.93

The difference in reported performance of the models occurs due to the specific images in the test set, since the test has only 1054 images. No replication of experiments or statistical tests were carried out, nor were confidence intervals computed. Thus, the rankings observed should not be considered as evidence of statistically significant differences between the results of the individual models, but as a comparison of the results of the models under the experimental conditions applied.

Fig. 9 illustrates the performance of DenseNet121 model over unseen data. The model classified (1004) samples accurately from total samples (1054) with only 50

misclassifications. These metrics provide valuable insights into the performance of model.

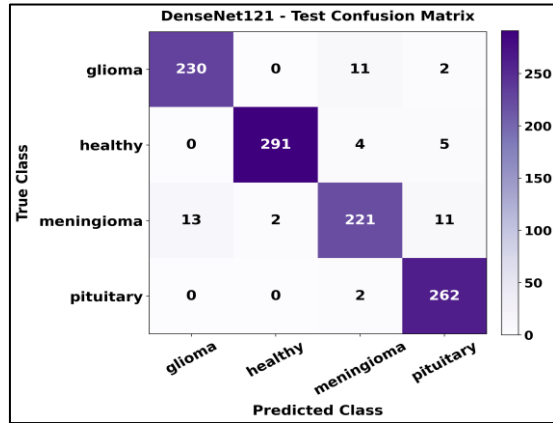


Fig. 9. Confusion matrix for DenseNet121.

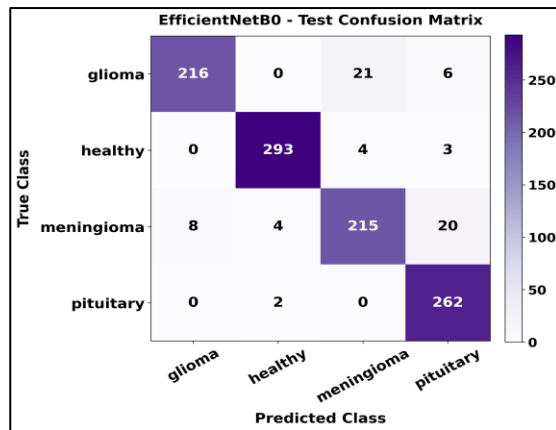


Fig. 10. Confusion matrix for EfficientNetB0.

Fig. 10 shows the performance of EfficientNetB0 model on test samples. (986) samples were correctly predicted out of total samples (1054) while 68 samples were wrongly predicted.

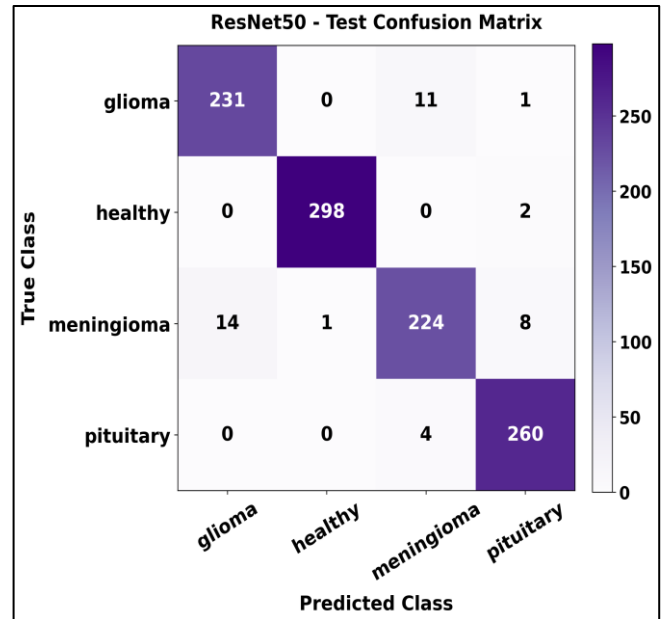


Fig. 11. Confusion matrix for ResNet50.

Fig. 11 illustrates the testing performance of the ResNet50 model. The model remarkably classified (1013) samples with only 41 misclassification predictions made by ResNet50. But, this is not enough to conclude that it is significantly better in performance compared to the other models. To establish consistency and reliability of the observed differences, repeated experiments, cross-validation, and appropriate statistical tests should be applied.

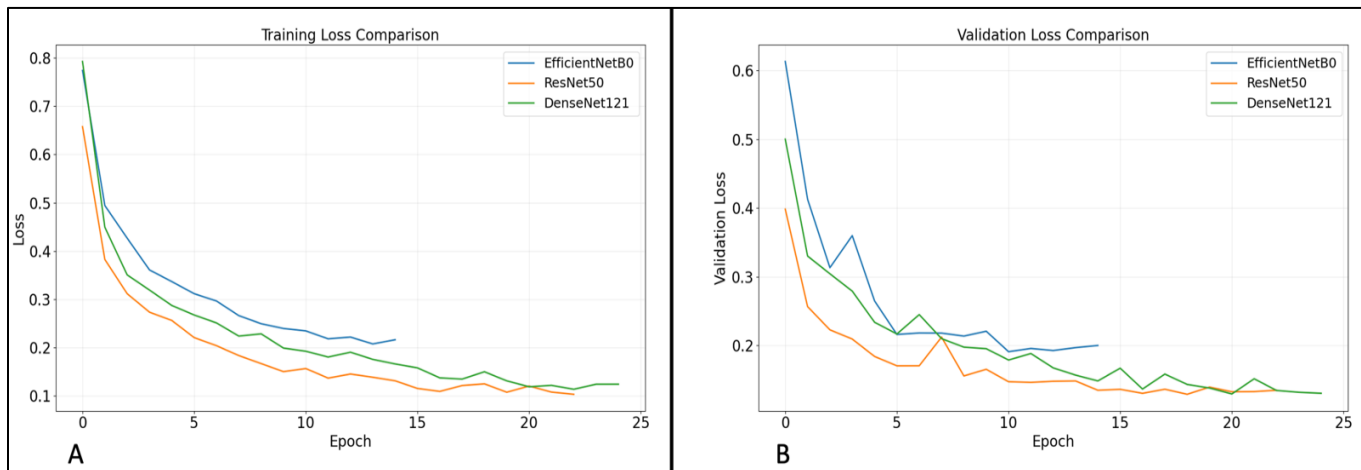


Fig. 12. Loss curves for all baseline models.

Fig. 12 represents the training and validation loss curves of all models. The blue curve shows the EfficientNetB0 while orange curve shows the ResNet50 and green curve shows the DenseNet121 model. In Fig. 12 (A), the training loss curves during the initial epochs of EfficientNetB0 and DenseNet121 overlap; however, the fluctuation pattern observed in the

DenseNet121 model. The ResNet50 model drops from a loss value of (0.5) and fluctuates towards increasing epochs. In contrast, Fig. 12 (B) shows the validation loss curves. A fluctuation pattern was observed among all models during training and overlaps at various epochs.

B. Monte Carlo Analysis

To quantify predictive uncertainty, (MC) Dropout is employed by performing multiple stochastic forward passes with dropout enabled during inference. Prediction confidence, predictive entropy and predictive variance are analyzed alongside classification performance to provide a more comprehensive assessment of model reliability prediction stability, and uncertainty across individual samples.

In this study, (T=20) was selected as the number of stochastic forward passes to estimate predictive uncertainty for each model using (training=True) during inference. The mean predictive probabilities across these passes were used to determine the final predicted class, while predictive entropy variance was calculated to characterize uncertainty. The same number was applied across all models to guarantee reliable evaluation settings. However, the use of 20 passes alone does not establish the convergence or stability of uncertainty estimates; therefore, this can be reflected as a limitation of the proposed study.

The MC Dropout results for the three tested models are presented in terms of predictive accuracy, confidence, entropy,

variance and high-confidence error rate in Table IV. The highest MC accuracy is achieved with ResNet50 (93%) along with a mean confidence of 0.91. EfficientNetB0, on the other hand, achieves the lowest error rate at high confidence, with just 0.18% of its predictions being erroneous at high confidence. The classification accuracy achieved by the deterministic inference and the MC Dropout evaluation differs because of the different prediction procedure adopted. Deterministic inference is based on only one forward pass, while MC Dropout uses 20 stochastic forward passes to calculate the average predictive probabilities. This averaging may change the predicted class for some test samples, leading to a difference in overall accuracy. The accuracy dropped from 95.3% (deterministic) to 93% (MC Dropout) for ResNet50. The same differences were found in EfficientNetB0 and DenseNet121. These results demonstrate that the accuracy is partially dependent on the inference procedure adopted. Thus, based on these results, it is important to use predictive entropy and variance for a more comprehensive evaluation of model uncertainty. But the differences in accuracy cannot account for the change in performance. Further investigation is needed for stochastic predictions and batch-normalization behaviour.

TABLE IV. MC DROPOUT COMPARISON

Models	MC Accuracy	Mean Confidence	Mean Predictive Entropy	Mean Predictive Variance	High-confidence Errors	High-confidence Errors (%)
ReseNet50	93%	0.91	0.22	0.006	7	0.66%
DenseNet121	91%	0.90	0.24	0.0067	8	0.75%
EfficientNetB0	90%	0.86	0.33	0.011	2	0.19%

Fig. 13 depicts the confidence distribution of EfficientNetB0 and ResNet50 models. In Fig. 13 (A), the confidence distribution with respect to correct and incorrect predictions of the EfficientNetB0 model is shown; the most accurate predictions are made at a range of confidence between about 0.9 and 1.0. Incorrect predictions are comparatively low and are mostly associated with lower confidence scores. This suggests that the model is generally confident when it identifies the image correctly, while its confidence tends to decrease for more challenging cases. On the other hand, Fig. 13 (B) depicts the ResNet50 model analysis, where the model exhibits a similar structure, with the correct predictions being more clustered

around 0.95–1.0 confidence. There are only a few mispredictions noticed. The narrower distribution suggests that ResNet50 is more confident in its predictions on the samples it was tested on. The high number of correct predictions at high confidence levels indicates that both networks have captured discriminative knowledge for the 5 classes. The relatively decreased confidence for incorrect predictions is also encouraging, as more challenging or ambiguous examples seem to result in less confidence. ResNet50 has a more focused high-confidence distribution compared to EfficientNetB0, suggesting greater reliability in prediction.

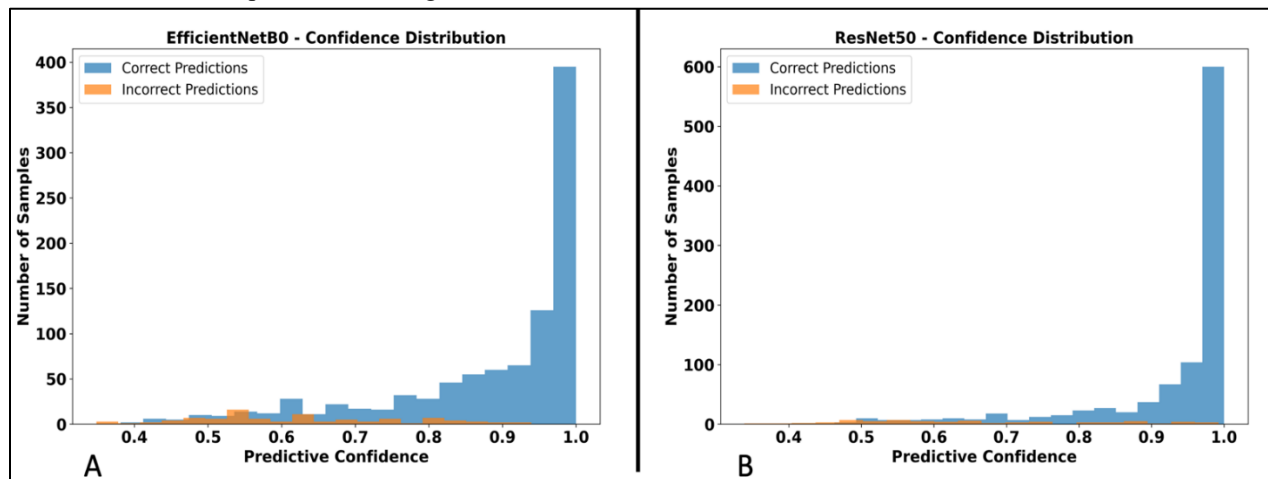


Fig. 13. Confidence distribution.

However, a high confidence score does not necessarily mean that the model is well calibrated. Expected Calibration Error (ECE) or reliability diagrams would be needed for a formal evaluation of calibration.

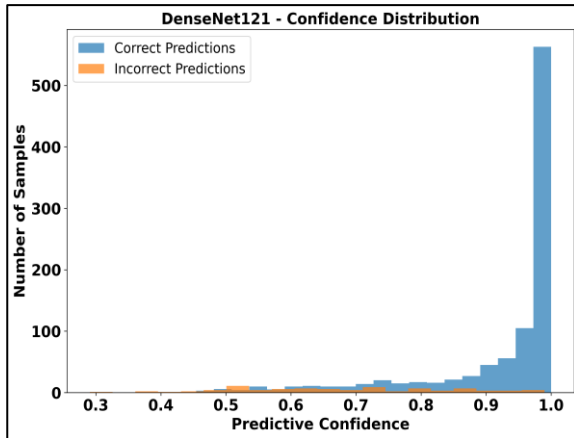


Fig. 14. Confidence distribution of the DenseNet121 model.

Fig. 14 represents the confidence distribution of the DenseNet121 model. The model is quite confident in its classification decisions, with the majority of the predictions falling within the 0.9–1.0 confidence range. Only a small number of incorrect predictions can be seen in the distribution. This behaviour indicates that DenseNet121 has acquired features that are useful for discriminating between the various classes and are able to provide accurate predictions for most of the test samples.

C. Uncertainty and Reliability

This section looks at the consistency of the confidence scores produced by the models when they are compared with their prediction performance. Aside from the basic classification results, probabilistic measures, entropy-based uncertainty, and calibration analysis are also included and deemed to give a more reliable understanding of the reliability of the model predictions.

The results of the reliability analysis are shown in Table V in terms of Brier score and negative log-likelihood (NLL). The

Brier score of 0.0597 reflects a relatively low level of error for the model's predictions of the probability of observations, and the NLL value of 0.113 suggests that the model's probabilities are appropriate. Generally, a lower value indicates better probabilistic performance for models evaluated under the same conditions. But these indicators do not provide a definitive answer as to the accuracy of the forecasts. The results should be interpreted in conjunction with the verification of the reported values and the confidence-bin analysis and reliability assessment.

TABLE V. PROBABILISTIC RELIABILITY METRICS

Probabilistic reliability metrics	Value
Brier Score	0.0597
NLL / log loss	0.113

TABLE VI. SUMMARY OF CONFIDENCE-BIN ANALYSIS

Bin	Samples	Mean Confidence	Accuracy	Gap
0.4-0.5	5	0.466	0.400	0.066
0.5-0.6	20	0.554	0.400	0.154
0.6-0.7	20	0.646	0.700	0.053
0.7-0.8	32	0.752	0.902	0.154
0.8-0.9	47	0.853	0.851	0.007
0.9-1.0	30	0.992	0.989	0.008

A summary of the confidence-bin analysis that was used for the examination of the calibration of the model is provided in Table VI. There are 930 samples in the range of 0.9–1.0, with most of the predictions falling within this range. The observed accuracy for this bin is 0.989, and the mean confidence is 0.992. The difference between the model's estimated confidence and the actual accuracy is only 0.0028, indicating that the model's confidence is well matched with how well it is performing in this high-confidence range. Some major differences were observed in the lower confidence intervals, particularly between 0.5 and 0.6 as well as 0.7 and 0.8. This change indicates that the model's calibration varies depending on its confidence level. Similarly, the overall ECE is low; calibration must be measured by the value in each bin and not by the value of the overall ECE.

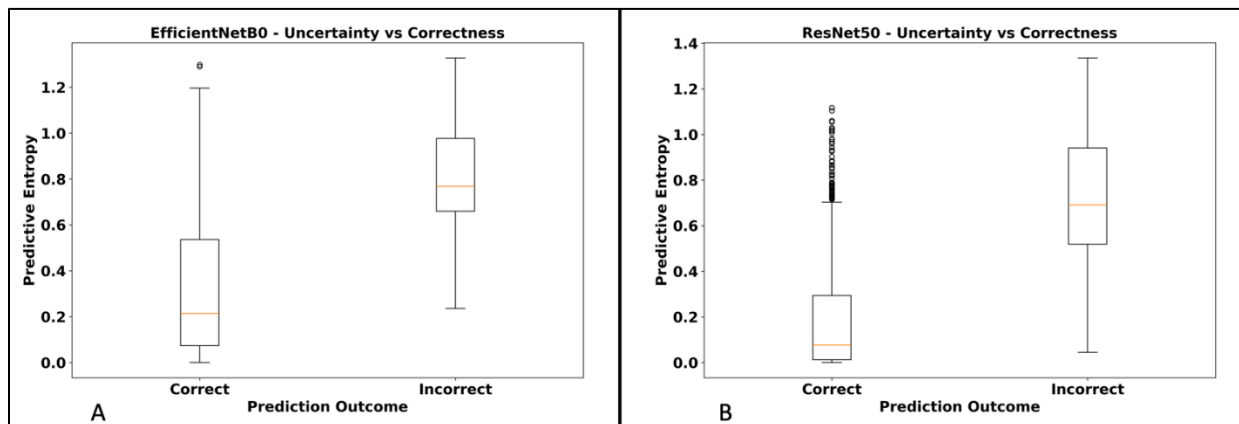


Fig. 15. Uncertainty vs correctness.

Fig. 15 illustrates the uncertainty vs correctness of EfficientNetB0 and ResNet50 models. In Fig. 15 (A), the

uncertainty and correctness of the EfficientNetB0 model are presented. Most of the samples that are correctly classified are

those with relatively low entropy, indicating that the model is fairly sure about its predictions. The misclassified samples, however, tend to have higher entropy, indicating that when the model misclassifies them, there is more uncertainty. This difference in the distribution is evident between these two groups. In Fig. 15 (B), the analysis of ResNet50 is shown. The lower entropy is correlated with correct predictions, and the higher entropy with incorrect predictions. However, a few samples are correctly classified as high-entropy outliers. From these cases, it is seen that the model can sometimes predict correctly even when the class-probability distribution of the model is relatively uncertain.

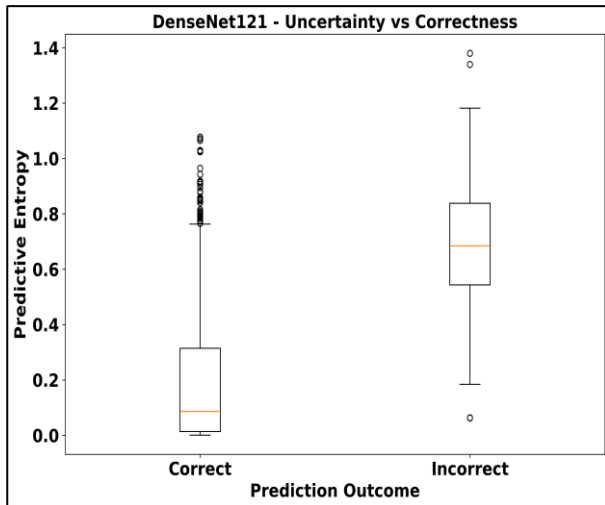


Fig. 16. Uncertainty vs Correctness for DenseNet121.

Predictive entropy represents the error or uncertainty the model has in predicting the class probabilities. If the entropy is high, then the probability will be spread out among the different classes; if low, then the prediction will be more confident. The greater entropy of the misclassifications, therefore, points to the tendency of both models to be less certain on challenging and ambiguous images. Overall, EfficientNetB0 achieves a more distinct separation between correct and incorrect samples, and the high entropy range of correctly classified samples with ResNet50 suggests that its uncertainty estimates are somewhat more variable.

The uncertainty analysis results of the DenseNet121 model are shown in Fig. 16. A lower entropy value would mean that the prediction is more confident, while a higher entropy would mean that the prediction is more uncertain, usually corresponding to unclear or even wrong classifications.

The reliability analysis results for the best overall model, ResNet50, are plotted in Fig. 17. The reliability curve reflects the closeness of the accuracy of the model's predictions to the actual accuracy at each level of confidence. The model achieved an EC of 0.012; this relatively low score indicates that confidence intervals among predicted classifications and model's self-reported accuracy. A model with a low ECE does not necessarily have a low error rate for all confidence intervals, though. The calibration of the model should therefore be checked from the diagram as well as from the Brier score, negative log-likelihood (NLL), and the confidence-bin statistics

in order to get a more comprehensive view of the model's calibration.

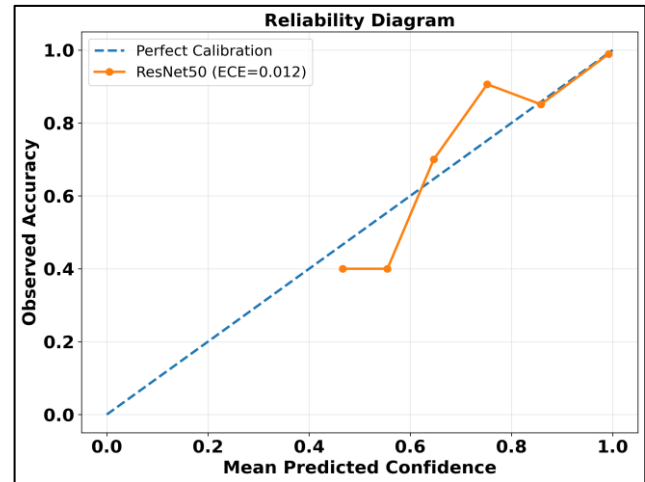


Fig. 17. Reliability of best performing model.

V. DISCUSSION

This study provides answers to the three research questions from complementary perspectives, namely, the quality of DFFs, their discriminatory power between classes of brain tumors and the value of uncertainty estimation for individual predictions. The results across all experiments indicate that ImageNet-pretrained CNNs can learn meaningful feature representations from brain MRI images; however, the resulting feature space exhibits limited class-wise clustering and separability. There is still substantial variation within each tumor class and overlap between the various classes in the learned feature space. Concurrently, the MC Dropout analysis suggests how uncertainty measures can provide valuable information concerning the reliability of model prediction, especially for challenging and ambiguous MRI cases.

These measures of confidence, entropy, variance, Brier score, and the NLL and ECE provide various viewpoints on model prediction and calibration. These, however, do not necessarily create clinical reliability or certify that estimates of uncertainty enhance diagnostic decisions. RQ2 is addressed at an exploratory level overall. While MC Dropout offers further information on predictive uncertainty, quantitative analysis and additional clinical validation are required to assess its utility in terms of detecting classification errors and making clinical decisions.

RQ1: Is the degree of representational learning of discriminative features better, worse or comparable in CNN models pre-trained on ImageNet for multi-class brain tumor classification?

The results show that CNN models pre-trained on the ImageNet dataset can learn useful features from brain MRI images, despite their training using natural images. The dataset consists of 7,023 MRI images, and EfficientNetB0 was able to extract 1280 features per image. The mean value of the extracted features and the mean feature norm are shown in Table I. The feature norms were also fairly similar in all four tumor classes seen in Fig. 4. This indicates that the magnitude of the extracted

features was comparable for the various categories of features. Further analysis of the distribution of the extracted features was carried out using (PCA). Based on Fig. 5 and Table I, the number of principal components required to explain 90, 95 and 99% of total variation were 371, 577 and 963, respectively. The results indicate that a large proportion difference in the extracted features captured by fewer dimensions than the original 1280. The variance explained by PCA does not necessarily reflect the robustness of the features to separate the classes of tumors.

Overall, the results of the feature analysis indicate that ImageNet-pretrained CNNs are able to learn useful representations from brain MRI images. The detailed analysis of features, however, was primarily performed with EfficientNetB0, and therefore cannot be directly compared across the three architectures. It is thus difficult to observe there were more discriminative features learned by one model than the other or if feature representations were similar across models. The results of the feature analysis did not yield any conclusion as to which architecture learned better representations, but differences were noted in the classification performance. ResNet50 had the highest overall performance of all the networks with 96% accuracy, 95% precision, 93% recall, 93% F1-score and an MCC of 0.94, as shown in Table III. The confusions matrices also indicate that ResNet50 successfully classified 1,013 test images against 1,004 images for DenseNet121 and 986 images for EfficientNetB0 as illustrated in Fig. 11, 9, and 10, respectively. The results show that ResNet50 achieved better classification accuracy under the conditions of this experiment. But the fact that it achieved higher classification accuracy does not necessarily imply that it learned better feature representations than the other two models.

RQ-2: Can MC Dropout-based uncertainty estimation be used to give meaningful insight into low confidence or ambiguous brain MRI predictions?

The results indicate that MC Dropout can offer additional insights into individual predictions beyond traditional classification metrics in the field of prediction. The predictive entropy of the incorrectly classified images was also higher with the evaluated models, but some correctly classified images had high entropy as well. This pattern suggests that the test set contains classification mistakes related to the impossibility of predicting the label. This is a clue that there is a possible connection between the difficulty of predicting the label and classification mistakes in the test set. This relationship was only explored using scatter plots, though. It is unclear if the ability of entropy to distinguish incorrect predictions from correct ones has been confirmed since AUROC analysis and risk-coverage/rejection curves were not conducted.

Therefore, RQ 2 is addressed at an exploratory level, as MC Dropout provides potentially useful uncertainty information, but its effectiveness in detecting unreliable predictions requires further quantitative evaluation.

1) *Practical implication:* The results of this study demonstrate that assessing the performance of classifiers designed for brain MRI classification using CNNs requires more than just classification accuracy. The learned feature representations, predictive uncertainty and the calibration of

predicted probabilities must also be taken into account. The observed correlation between the increase of the predictive entropy and the error of prediction proposes to investigate the behavior of the models by using uncertainty measures. This study, however, has not quantitatively determined the reliability of these measures in determining classification errors. Similarly, the calibration results represent the quality of predictions on the assessed dataset and do not measure clinical effectiveness. Future studies should involve external validation, quantitative analysis of the uncertainty-based error detection, and testing in clinical workflows to assess the potential benefits of these techniques for clinical diagnosis.

2) *Limitations:* The results must be interpreted with several Limitations. The first limitation of the proposed study is a single publicly available brain MRI dataset with four classes. This dataset will help compare models, but may not represent the full spectrum of patient populations, MRI imaging protocols, scanners, or hospitals. Moreover, the dataset does not contain patient-level data, which makes difficult to confirm the data spilt at patient-level. This dataset will be helpful for comparing models, but may not represent the full spectrum of patient populations, MRI imaging protocols, scanners, or hospitals. Secondly, the relationship between predictive entropy and misclassification was investigated via scatter plots. The validity of the ability of uncertainty measures to distinguish between correct predictions and errors, however, was not quantitatively verified by AUROC analysis or risk-coverage or rejection curves. Third, there was no evaluation of the ability of uncertainty-based case prioritization to work in a clinical context or whether it would aid in reducing diagnostic error and improve diagnostic decisions. Fourth, MC Dropout was used, with 20 stochastic forward passes being used, though no experiments were run to see if this number of forward passes was enough to provide stable uncertainty estimates. Using 'training=True' at inference time could also impact the behaviour of batch-norm, depending on model implementation, which needs to be explored further.

VI. CONCLUSION

In this work, we explore how the features learned by ImageNet-trained CNNs are used for brain MRI classification and if the prediction uncertainty of such models can be used to gain insight into the classification beyond traditional metrics. The analysis revealed that the pretrained networks were able to learn meaningful representations of the data, though the learned feature space did not reflect distinct clusters of tumors. The original 1,280-dimensional representation can be summarized into just 371, 577, and 963 principal components that would represent 90%, 95%, and 99% of the total variance. Concurrently, the between-to-within-class separability ratio was 0.148, indicating that there is a great deal of overlap between classes and a lot of variation within classes. This points out that if the tumor categories are naturally well separated, it is not necessarily so if they are also naturally well-separated in terms of informative deep features. The ResNet50 CNN achieved the best classification performance with an accuracy of 96%, a precision of 95%, a recall of 93%, an F1-score of 93%, and an

MCC of 0.94, among the evaluated CNNs. The confusion matrix had 1013 images correctly classified from 1054 images that were tested. These findings demonstrate that useful classification boundaries can be learned even though the feature distributions of the classes have significant overlap. The MC Dropout experiments provided an additional valuable perspective. It was found that classification accuracy alone is not enough as an indicator of the reliability of a model's individual predictions. The uncertainty was typically higher for incorrect predictions, which were associated with higher predictive entropy, implying that there was greater uncertainty when the model had a more difficult task to differentiate classes. The mean prediction confidence for ResNet50 was 0.91 and the variance of the predictions was 0.006. An ECE of 0.012 was obtained in the reliability analysis, which suggested that the model's confidence level was in good agreement with its observed performance.

The findings indicate that the CNN models studied may be applied in an automatic brain MRI classification system for the conditions investigated in this study. Predictive uncertainty can be used as further assistance by identifying instances that should be looked at in detail. For example, in a future HITL system, the predictions with high entropy could be submitted for a radiologist to verify. This study did not, however, explore referral thresholds, radiologists' decisions, diagnostic findings, or impact on clinical workflow efficiency. The usefulness of uncertainty-based case prioritization for the use of clinical guidelines is thus yet to be validated. It is important to note that the models should not replace expert clinical judgment, but should be used as research tools for automated classification.

FUTURE WORK

Several avenues are available to further develop the results of this study. First, the learned representations need to be fine-tuned to the particular task, refined with attention to the MRI data, and learned with the metric-learning methods. The procedures may contribute to a reduction of feature overlap and enable a better separation of the tumor classes. Further experiments might also explore the possibility of obtaining similar uncertainty estimates using MC Dropout as compared to ensemble-based methods and of further refining the accuracy of the model's confidence using other calibration methods.

DATA AVAILABILITY STATEMENT

The dataset used in this study for detection of cyberbullying classification is publicly available at <https://www.kaggle.com/datasets/rm1000/brain-tumor-mri-scans>.

REFERENCE

- [1] A. A. Asiri, T. A. Soomro, A. A. Shah, G. Pogrebna, M. Irfan, and S. Alqahtani, "Optimized Brain Tumor Detection: A Dual-Module Approach for MRI Image Enhancement and Tumor Classification," *IEEE Access*, vol. 12, pp. 42868–42887, 2024, doi: 10.1109/ACCESS.2024.3379136.
- [2] U. Raghavendra et al., "Brain tumor detection and screening using artificial intelligence techniques: Current trends and future perspectives," *Computers in Biology and Medicine*, vol. 163, p. 107063, Sep. 2023, doi: 10.1016/j.combiomed.2023.107063.
- [3] R. Haque et al., "Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis," *Computers in Biology and Medicine*, vol. 191, p. 110166, Jun. 2025, doi: 10.1016/j.combiomed.2025.110166.
- [4] M. A. Alam, A. Sohel, K. M. Hasan, and I. Ahmad, "Advancing Brain Tumor Detection Using Machine Learning And Artificial Intelligence: A Systematic Literature Review Of Predictive Models And Diagnostic Accuracy," *Strategic Data Management and Innovation*, vol. 1, no. 01, pp. 37–55, Dec. 2024, doi: 10.71292/sdmi.v1i01.6.
- [5] A. Batool and Y.-C. Byun, "A lightweight multi-path convolutional neural network architecture using optimal features selection for multiclass classification of brain tumor using magnetic resonance images," *Results in Engineering*, vol. 25, p. 104327, Mar. 2025, doi: 10.1016/j.rineng.2025.104327.
- [6] S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi, and M. Abedalthagafi, "Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment," *npj Precis. Onc.*, vol. 8, no. 1, p. 80, Mar. 2024, doi: 10.1038/s41698-024-00575-0.
- [7] Md. Ariful Islam, M. F. Mridha, M. Safran, S. Alfarhood, and Md. Mohsin Kabir, "Revolutionizing Brain Tumor Detection Using Explainable AI in MRI Images," *NMR in Biomedicine*, vol. 38, no. 3, p. e70001, 2025, doi: 10.1002/nbm.70001.
- [8] A. Rahman, M. Hayat, N. Iqbal, F. K. Alarfaj, S. Alkhalaf, and F. Alturise, "Enhanced MRI brain tumor detection using deep learning in conjunction with explainable AI SHAP based diverse and multi feature analysis," *Sci Rep*, vol. 15, no. 1, p. 29411, Aug. 2025, doi: 10.1038/s41598-025-14901-4.
- [9] S. Das and R. S. Goswami, "Advancements in brain tumor analysis: a comprehensive review of machine learning, hybrid deep learning, and transfer learning approaches for MRI-based classification and segmentation," *Multimed Tools Appl*, vol. 84, no. 23, pp. 26645–26682, Jul. 2025, doi: 10.1007/s11042-024-20203-0.
- [10] T.-W. Wang et al., "Artificial Intelligence Detection and Segmentation Models: A Systematic Review and Meta-Analysis of Brain Tumors in Magnetic Resonance Imaging," *Mayo Clinic Proceedings: Digital Health*, vol. 2, no. 1, pp. 75–91, Mar. 2024, doi: 10.1016/j.mcpdig.2024.01.002.
- [11] B. M. Varna, P. Rakesh, V. Hemanth, B. Mukesh, and A. J., "Uncertainty-Aware Brain Tumor Segmentation in MRI: A Res-UNet Framework with Monte Carlo Dropout," in 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Jul. 2025, pp. 1–9, doi: 10.1109/ICDICI66477.2025.11135391.
- [12] G. E. Mosquera Rojas, S. van der Voort, C. M. Pirkil, S. Kaushik, M. Smits, and S. Klein, "Evaluation of Monte Carlo Dropout for Uncertainty Quantification in Multi-task Deep Learning-Based Glioma Subtyping," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, C. H. Sudre, M. I. Hoque, R. Mehta, C. Ouyang, C. Qin, M. Rakic, and W. M. Wells, Eds., Cham: Springer Nature Switzerland, 2026, pp. 180–190, doi: 10.1007/978-3-032-06593-3_17.
- [13] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation," *Front. Neurosci.*, vol. 14, Apr. 2020, doi: 10.3389/fnins.2020.00282.
- [14] B. Shaheema, N. B. Muppalaneni, and K. S. Devi, "An explainable deep learning-based panoptic segmentation for brain tumor diagnosis," *Neural Comput & Applic*, vol. 37, no. 25, pp. 20639–20662, Sep. 2025, doi: 10.1007/s00521-025-11459-0.
- [15] T. Zhou, S. Ruan, and B. Lei, "BUFNet: Boundary-aware and uncertainty-driven multi-modal fusion network for MR brain tumor segmentation," *Medical Image Analysis*, vol. 107, p. 103855, Jan. 2026, doi: 10.1016/j.media.2025.103855.
- [16] N. Rasool and J. I. Bhat, "Brain tumour detection using machine and deep learning: a systematic review," *Multimed Tools Appl*, vol. 84, no. 13, pp. 11551–11604, Apr. 2025, doi: 10.1007/s11042-024-19333-2.
- [17] S. Mohsen, S. Oraby, and M. Abdel-Aziz, "Deep Learning and Machine Learning for Brain Tumor Detection: A Review, Challenges, and Future Directions," *Arch Computat Methods Eng*, vol. 33, no. 3, pp. 3931–3955, Apr. 2026, doi: 10.1007/s11831-025-10416-3.
- [18] N. M. Hussain Hassan and W. Boulila, "Efficient Approach for Brain Tumor Detection and Classification Using Fuzzy Thresholding and Deep Learning Algorithms," *IEEE Access*, vol. 13, pp. 78808–78832, 2025, doi: 10.1109/ACCESS.2025.3566332.

- [19] M. Nahiduzzaman et al., "A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images," *Sci Rep*, vol. 15, no. 1, p. 1649, Jan. 2025, doi: 10.1038/s41598-025-85874-7.
- [20] R. Disci, F. Gurcan, and A. Soylu, "Advanced Brain Tumor Classification in MR Images Using Transfer Learning and Pre-Trained Deep CNN Models," *Cancers*, vol. 17, no. 1, p. 121, Jan. 2025, doi: 10.3390/cancers17010121.
- [21] "Brain Tumor (MRI Scans)." Accessed: Sep. 18, 2026. [Online]. Available: <https://www.kaggle.com/datasets/rm1000/brain-tumor-mri-scans>
- [22] O. A. M. F. Alnaggar et al., "Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis," *Artif Intell Rev*, vol. 57, no. 8, p. 221, Jul. 2024, doi: 10.1007/s10462-024-10814-2.
- [23] T. Islam, Md. S. Hafiz, J. R. Jim, Md. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, p. 100340, Jun. 2024, doi: 10.1016/j.health.2024.100340.
- [24] P. Dehbozorgi, O. Ryabchykov, and T. W. Bocklitz, "A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities," *Computers in Biology and Medicine*, vol. 187, p. 109768, Mar. 2025, doi: 10.1016/j.combiomed.2025.109768.
- [25] M. T r, M. Gupta, A. T a, V. Kumar V, O. Geman, and D. Kumar V, "An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging," *Journal of Neuroscience Methods*, vol. 410, p. 110227, Oct. 2024, doi: 10.1016/j.jneumeth.2024.110227.
- [26] M. M. M, M. T. R, V. K. V, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50," *BMC Med Imaging*, vol. 24, no. 1, p. 107, May 2024, doi: 10.1186/s12880-024-01292-7.
- [27] M. Rasheed et al., "Improved brain tumor classification through DenseNet121 based transfer learning," *Discov Onc*, vol. 16, no. 1, p. 1645, Aug. 2025, doi: 10.1007/s12672-025-03501-3.