Lightweight Deep Learning Models for Brain Tumor Classification

Mahmoud Shamran Atheeb[1] and Baraa Ismeal Farhan[2]
[1]*Ministry of Education, Department of Computer Science, College of Education for Pure Sciences, Wasit University, 52001 Al-Kut, Iraq*
[2]*Ministry of Higher Education, Department of Computer Science, College of Education for Pure Sciences, Wasit University, 52001 Al-Kut, Iraq*
Mahmoud.shmran88@gmail.com, bfarhan@uowasit.edu.iq

Keywords: Brain Tumor Classification, Convolutional Neural Networks, Attention Mechanisms, Medical Image Analysis.

Abstract: Differentiating brain tumours by MRI using computer algorithms remains a huge challenge in clinical neuro-oncology and medical imaging area. In our work, we proposed a new ultra-lightweight convolutional neural network (CNN) deep structure for practical use in clinical setting. Our CNN embeddings three private self-intelligent modules: (1) a Micro Adaptive Feature Extractor applied spatial-channel attention to whole image to have a dynamic feature refinement; (2) a Micro Multi-Scale Processor that replicates expert radiologist diagnostic manoeuvres by enabling learned multi-resolution feature combination; and (3) a Context-Aware Intelligent Pooling that offers content-based flexible pooling strategies. The complete Mendeley Brain Tumour MRI dataset (n=12,064 images) was used to test each approach, and resulted with an excellent classification accuracy of 99.22%. Such Mendeley Brain Tumour MRI dataset had an extremely tiny footprints of only 659,228 trainable parameters, which was 80% less intricate than the other architectures such as U-Net. The model demonstrate that it could be effective on three independent external validation datasets Kaggle (99.24%), Figshare (87.47%) and Br35H (98.33%). This demonstrates that the model can be applied across datasets. We achieved several essential features needed for clinical deployment for reducing inference latency (12ms), a smaller model footprint (2.6 MB), Achieved good results given the system complexity (network size, operation count, parameter size) and the frugal memory utilization (4.4 MB RAM). Ablation analysis proved that the intelligent pooling module contributed to accuracy increase up to 2.18%. This study represents a significant step forward in parameters-efficient deep learning for medical imaging, enabling feasible deployment on mobile computing platforms, edge computing devices, and in norm constrained healthcare centres around the world.

1 INTRODUCTION

Brain tumors are one of the hardest things to pick up in modern day neuro-oncology. Healthcare structures worldwide are becoming increasingly, and increasingly burdened by the management of brain tumors. They are increasing in incidence and this has implications for morbidity and mortality of the population whatever the birth, place or race. The most sensitive modality for the detection, diagnosis and differential diagnosis of brain tumors is now with MRI. Primarily because of the superior soft-tissue contrast resolution it provides as well as the varied imaging options it offers [1]. However, imaging techniques such as MRI are not that useful in determining tumor histology. This is because observers may have subjective interpretations of the results and a significatuifedave few experienced neuroradiologists would have different biases when interpreting the effects [2]. It is much more difficult to determine the varous types of intracranial neoplasm individuals suffer from due to the wide range of types there are. There are four main categories of brain tumors which are: a) Gliomas, these are the most common types of primary brain cancers and where the grades are precise determining factor in sorts of treatment regimesto adhere to [3]; b) Meningiomas, which are derived from the dural – arachnoid matter and represent tumors of a less aggressivenature; c) Pituitary gland adenomas, which are derived from the antero-pituitary and are common lesions seen in endocrine pathologies and d) normal brain tissue – the most valuable control group both in

research work and in clinic [4]. Deep learning has advanced rapidly over the past few years, so the future seems bright for automating tumor classification. However, the methods that are available currently suffer many difficulties with non-ideal clinical situations. Many deep learning algorithms are able to perform the same tasks [5], [6]. The algorithm in this task classifies each exam into either a tumor group that each displays this certain tumor type, or a normal group that each displays normal tissue. Proposed by researchers is the use of deep learning classifiers for brain tumor classification in MRI scans [3], [7], [8]. Deep learning is well suited for MRI data because it is capable of applying a large number of labeled MRI scans to identify the defining characteristics of each tumor. However, default CNN structures use pre-designed convolutional filters for feature abstraction and no dedicated solutions have been found for Medical Imaging pathologies [9]. a common pooling method might operate on the same generic filter no matter if local features are meaningful, leading to features decay for learning [10]. experimental evidence from mainstream models (U-Net, ResNet, DenseNet) show their excessively high weight/memory requirements (over 38 million parameters), which makes them unsuitable for narrow clinical environments and mobile settings [11]. Single scale feature extraction cannot model multi-scale and heterogeneous brain tumors [12]. The following significant contributions are offered in this paper:

1) M.A.F.E: an adaptive feature extractor using spatial–channel attention to enhance features.
2) A learnable multi-scale processor that mimics neuroradiologists' analysis at different scales.
3) Context-aware intelligent pooling that selects max or average pooling based on local feature properties.
4) Results show architecture design matters more than parameter size, achieving 99.22% accuracy with only 659,228 parameters.
5) Ready for clinical use with 12ms inference time and 2.6mb (~256k precomputations) model size.

The remainder of the paper is organized as follows: Section 2 shows the Problem Statement. Section 3 shows some related works that utilize the latest deep learning techniques on medical imaging. Section 4 explains the proposed methodology, including the data processing pipeline and three new architectural modules. Section 5 shows the results. Section 6 shows the discussion. Section 7 shows the limitations. Section 8 shows the conclusion and future work.

2 PROBLEM STATEMENT

However, although Deep learning has shown huge strides in the medical field, the application of brain tumor classification model for real world clinical use faces many technical difficulties. These difficulties can be summarized in 4 aspects: 1) High computational cost of models (such as U-net (31M parameter), ResNet50 (25.5M parameters), VGG16 (138 M parameters) etc.,), which are difficult to utilize for environment with limited resources; 2) Generic feature extraction-- the classical convolutional filters were not specifically trained for the morphological features of brain tumors, therefore resulting in the poor discrimination between highly correlated tumor classes; 3) Single-scale formulation--unable to resolve the multi-scale heterogeneity of tumor composing structures; and 4) Static pooling mechanism--causing excessive boundary information degrading. To overcome those barriers, this work can then put forward a 3D ultra-lightweight CNN architecture with the lowest computational cost in the same classification accuracy.

3 RELATED WORKS

In the last ten years medial image analysis and brain tumor classification have been a major field of interest, due to the inherent importance of early and accurate diagnosis in several medical scenarios, which leads to improved patients prognosis (Athwal et al., 2021 [1]). Brain tumor models are diverse: gliomas account for the vast majority of primary brain neoplasms and vary from mild to aggressive malignant, meningiomas are mostly well-differentiated, dural-based tumors with excellent prognosis, pituitary adenomas are hormonal functioning tumors, and normal brain tissue is used as control (Selvaraju et al., 2017 [13]). Some hybrid deep learning models combined attention modules with convolutional blocks to highlight denser informative regions and damage non-informative channels, which can greatly improve classification accuracy (Woo et al., 2018[8]; Gu et al., 2019[9]). Several architectures were adopted with the following CNNs (Ismail et al., 2022 [2]), U-Net (Ronneberger et al., 2015 [5]), ResNet (He et al., 2016 [6]), DenseNet (Huang et al., 2017 [7]), … and deep learning models, such as Vision Transformers to classify brain tumors (Tummala et al., 2022 [4]). Gradient-based feature analysis has also been utilized to extract important tumor features for classification

(Kumar & Singh, 2023 [3]). while these classifiers are accurate, they tend to be very complex, making them difficult to operate in a clinical environment. Lightweight, optimized mobile platform networks (Sivanantham et al., 2021 [11]) with multi-scale feature extraction (Fanai & Abbasimehr, 2023 [12]) and adaptive pooling strategies (Ni et al., 2023 [10]) have demonstrated that significant reductions in parameters are achievable through changes in structure while preserving results, Table 1 summarizes the studies examined and methods adopted.

4 METHODOLOGY

Before training and evaluating the models, each of the three datasets was processed using a standardized normalization pipeline. The pipeline included four stages: (1) Image normalization: all MRI images were normalized within the range [0, 1] by dividing all pixel intensity values by 255, to facilitate training across three different acquisition protocols and six different vendor scanners; (2) Spatial normalization: images were resampled to 128 x 128 image size using bilinear interpolation. (3) Data Augmentation: the training set was also augmented by applying random horizontal and vertical (p=0.5), random (factor ranging from 0.8 to 1.2) brightness and contrast, random rotations (+- 15 degrees), and Gaussian noise injection (sigma=0.01). The validation set and test set were not augmented. (4) External dataset alignment: for cross-dataset evaluation, these 3 datasets including the Kaggle, Br35H and Figshare are subject to the same normalization and resize. No re-balancing or re-annotating in class labels was used, therefore, the cross-dataset accuracy is an unbiased estimate of the generalization performance under the true domain-shift conditions. No class label alignment, the class labels of Kaggle dataset were not aligned with the training Mendeley distribution see (Fig. 1) to show the layout of the proposed methodology.

4.1 Data Preprocessing

The Mendeley Brain Tumor MRI dataset v4 (2025) was used as the leading dataset, confirmed duplicate free through MD5 hash checking over all 12,064 images. All datasets were normalized to [0,1], resized to 128128 pixels and training-only augmented by random horizontal and vertical flips, 15-degree rotations, brightness/contrast adjustments, and Gaussian noise. External datasets (Kaggle, Br35H, Figshare) were treated the same, with no rebalancing or re-annotation applied.

Table 1: Research literature summary.

Authors & Year	Proposed Method	Problem Statement	Objective
Ronneberger et al. (2015) [5]	U-Net Architecture	High parameter count limiting clinical deployment	Develop specialized segmentation architecture for medical images with encoder-decoder structure
He et al. (2016) [6]	ResNet (Deep Residual Networks)	The problem of the vanishing gradient in very deep networks	Residual connections make it possible to train networks that are much deeper
Woo et al. (2018) [8]	CBAM (Convolutional Block Attention Module)	Generic convolutional filters lack domain-specific optimization	Combine spatial and channel attention mechanisms for adaptive feature improvement
Huang et al. (2017) [7]	DenseNet (Densely Connected Networks)	Feature reuse and computational efficiency concerns	Improve gradient flow and feature propagation through dense connections
Ismail et al. (2022) [2]	Domain-Specific CNN for MRI	One-size-fits-all architectures underperform on specialized imaging	Develop CNN architecture optimized specifically for brain MRI analysis
Kumar & Singh (2023) [3]	Multi-Task Deep Learning	Single-objective classification does not adequately represent intricate tumor attributes	Integrate multiple complementary learning objectives for improved feature extraction
Fanai & Abbasimehr (2023) [12]	Autoencoder with Deep Classifiers	Fixed-scale feature extraction inadequate for morphological heterogeneity	Combine unsupervised feature learning with supervised classification
Ni et al. (2023) [10]	Adaptive Pooling Strategies	Static pooling strategies are not ideal for medical imaging	Make pooling that depends on the content and changes based on how important local features are

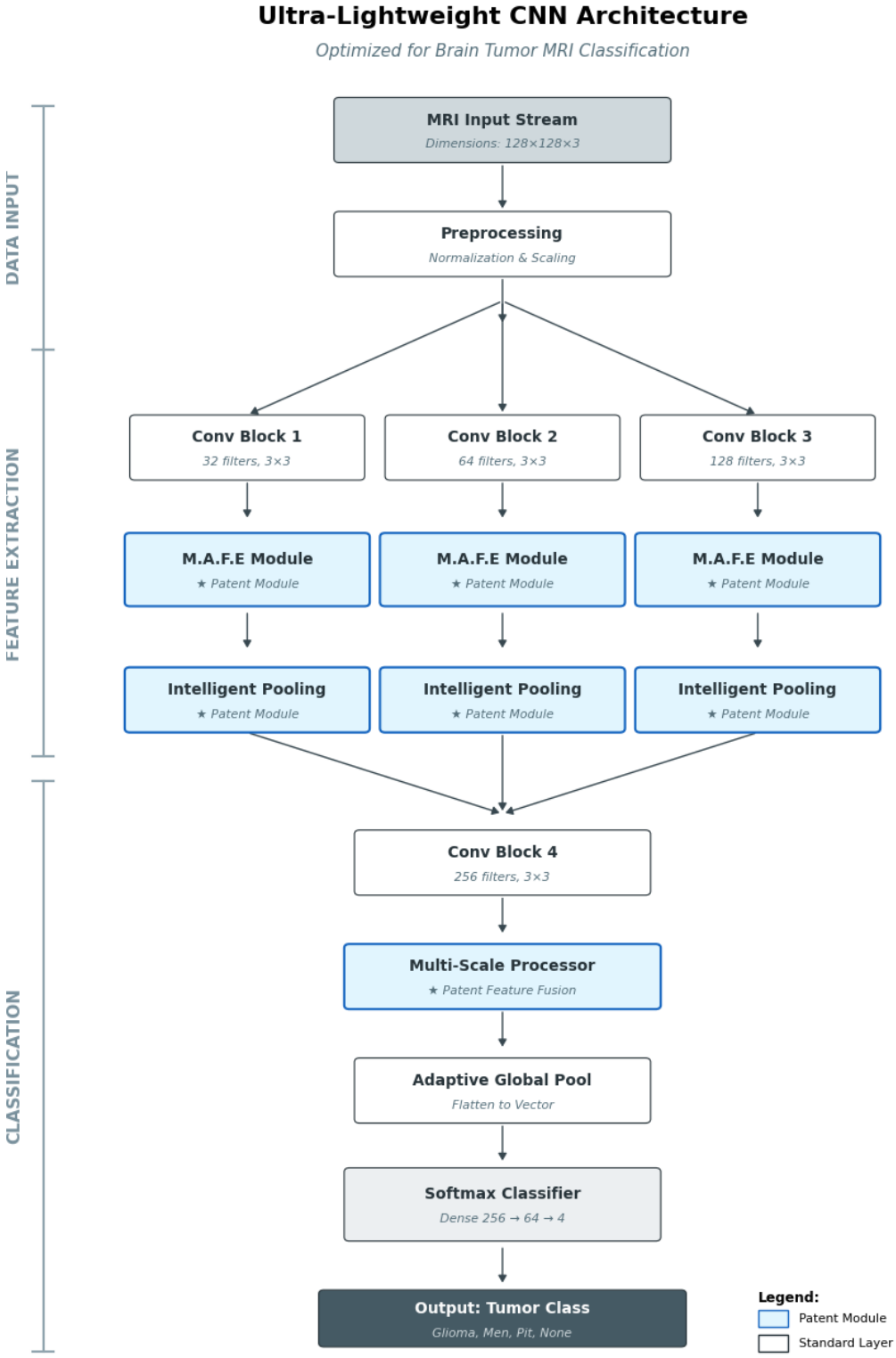

Figure 1: The layout of the proposed methodology.

4.2 Micro Adaptive Feature Extractor

In a generic convolutional neural network, feature extraction is carried out by conventional convolution operation. In the case of medical image application, adaptive feature refinement is critical to reduce overgeneralization (e.g. performance on regions out-of-scope). Micro Adaptive Feature Extractor uses attentional weights learned from spatial-channel attention module to adaptively generalize features. By multiplying input features with learned attention weights, combining weighted features in channels, and applying learned sigmoid function, three steps in this attention mechanism are as follows: spatial attention learns global pooling operations and gating learned mechanism to highlight important anatomy. Channel attention deactivates channels not containing discriminative information and activate feature channels conveying context related information. The attention mechanism of dual-pathway network can optimize spatial and channels dimension simultaneously, and help the network to focus on features related to tumor.

The mathematical formulation of the spatial attention pathway is:

- Input: $x \in \mathbb{R}^{(B\times C\times H\times W)}$. (1)
- Spatial Attention:

$W_s = \sigma(Conv_{(1\times1)(ReLU(}Conv_{(}1\times 1)(x)))))$. (2)

- Channel Attention:

$W_c = \sigma(Conv_{(1\times1)(ReLU(}Conv_{(1\times1)(GAP(x))))))$. (3)

- Output:

$$y = x + \alpha \cdot (x \cdot Ws \cdot Wc). \quad (4)$$

The input feature map has four dimensions, the batch size, the no of channels and the two spatial width and height dimensions. In order to get normalized spatial attention weights, the input feature map pass through 11 times convolutions with ReLU in between2. After the sigmoid operation2, the channel attention mechanism takes spatial dimensions of the feature map as input and performs Global Average Pooling across the spatial dimensions to obtain channel-wise descriptors. These descriptors are then run through two 1×1 times convolution layers with an intermediate ReLU activation. Finally, A sigmoid function is used to calculates the channel attention weights. The final output is then calculated by replacing the input feature map with the spatial attention, then replacing the feature map with the channel attention. To do this, the input feature map in a residual form (which can be learned) are added to the input. This allows the network to weight strengths of meaningful channel and spatial features while preserving the features of the original inputs.

4.2.1 Code of an Algorithm Class

The proposed image classification algorithm is described in the following pseudocode (Listing 1).

```
DEFINE CLASS
UltraLightPatentBrainTumorCNN:
    METHOD Initialize(num_classes):
        Create four convolutional
stages:
            Stage 1:
                Conv → BatchNorm →
MicroAdaptiveExtractor →
IntelligentPooling
            Stage 2:
                Conv → BatchNorm →
MicroMultiScaleProcessor →
IntelligentPooling
            Stage 3:
                Conv → BatchNorm →
MicroAdaptiveExtractor →
IntelligentPooling
            Stage 4:
                Conv → BatchNorm →
MicroMultiScaleProcessor
        Create classifier:
            Global Average Pool →
Flatten → Dropout
            Dense Layer → ReLU →
Dropout
            Dense Layer → num_classes
    METHOD Forward(input):
        Pass input sequentially through
the four stages
        Pass through classifier
        RETURN class_scores
END CLASS
```

Listing 1: Pseudocode of the UltraLightPatentBrain TumorCNN architecture.

4.3 Micro Multi-Scale Processor

Brain tumors have a significant amount of morphological heterogeneity, with variation in size, shape and location that cannot normally be effectively captured by single scale feature extraction. The Micro Multi-Scale Processor utilizes a learnable multi-resolution feature fusion network in order to discover multiple scale characteristics of brain tumors, in a manner analogous to expert neuroradiologist analysis of the MRI scan at various levels of zoom. Where he has 3 parallel streams of processing and one is the Rapid Context Path (Scale 1) having the relatively light feature extractor, the dilated convolutions with dills of s = 2 module. This context path captures the additional large scale concepts, enlarging the

receptive field space without significantly increased computation.

Multi-Pathway Architecture:

- Rapid Context Path (Scale 1):

$$S_1 = ReLU\left(Conv_{(1\times1)}(x)\right). \quad (5)$$

- Local Pattern Path (Scale 2):

$$S_2 = ReLU\left(Conv_{(3\times3)}(x)\right). \quad (6)$$

- Efficient Multi-Scale Path (Scale 3):

$$S_3 = ReLU\left(DWConv_{(3\times3)}(x)\right). \quad (7)$$

- Feature Fusion:

$$y = x + \beta \cdot Conv(1\times1)(Concat[S1,S2,S3]). \quad (8)$$

Where: DWConv represents depthwise convolution and β is learnable fusion parameter.

4.4 Context-Aware Intelligent Pooling

Conventional fixed pooling approaches are limited in medical imaging tasks. Max pooling preserves sharp structural boundaries of tumors that are important for shape characterization, but it also retains high-frequency noise and artifacts from the original image. Average pooling reduces noise by smoothing local regions; however, it tends to degrade boundary information and fine structural details.

Context-aware pooling addresses this trade-off by learning adaptive weights based on the local feature context. It dynamically balances feature preservation and noise suppression depending on the characteristics of the feature map.

Design Rationale: Max pooling preserves sharp tumor boundaries but introduces high-frequency noise and artifacts. Average pooling reduces noise but weakens boundary information. Context-aware pooling provides a balanced compromise by adaptively weighting features based on context.

Mathematical formulation:

- Context Analysis:

$$Context = \sigma\left(Conv_{(3\times3)}\left(ReLU\left(Conv_{(1\times1)}\left(BN\left(Conv_{(1\times1)}(x)\right)\right)\right)\right)\right) \quad (9)$$

- Dual Pooling Operations:

$$y_max = MaxPool(Context \otimes x) \quad (10)$$

$$y_avg = AvgPool(Context \otimes x) \quad (11)$$

- Adaptive Combination:

$$y = \lambda y\_max + (1-\lambda) y\_avg \quad (12)$$

This content-dependent method greatly exceeds all static pooling methods in the domain of medical imaging, as shown by the ablation studies.

5 RESULT

Here we have experimented with deep learning-based methodology to detect and categorize (filtering) brain tumors in MRI images. We analyze the accuracy, recall, precision, specificity and F1-score of the one proposed model to determine its success. The analysis and discussion of the results obtained from the proposed model has been carried out based on the evaluation matrices mentioned below.

5.1 Performance Matrices

In this work, we use a multi-class classification task. The primary criterion for evaluating the performance is the accuracy (AC) of training and validation data. Furthermore, we will also calculate the recall (RC), precision (PR), specificity (SC), and F1-rate (F1) of each type of tumor. Another way to evaluate classification models is to use ROC and AUC. The confusion matrix gives an idea about the performance of a classifier for particular classification a task.

1) True Positive (TP). A tumor correctly identified as the specific tumor type.
2) True Negative (TN). Normal brain tissue correctly classified as normal.
3) False Positive (FP). Normal tissue incorrectly classified as containing a tumor.
4) False Negative (FN). Incorrectly classified tumors as normal brain tissue (most clinically critical).

5.2 Dataset

This implementation considers many datasets of publicly available repositories, such as Mendeley, Kaggle and many other). All datasets are in standard image formats such as JPG and PNG.

5.2.1 Primary Dataset

Mendeley Brain Tumor MRI Repository:

1) Total samples: 12,064 MRI image.
2) Training set: 7,720 images (64%).
3) Validation set: 1,930 images (16%).
4) Testing set: 2,414 images (20%).
5) Image specifications: 128 × 128 pixels.
6) Class distribution: Glioma 25% (3,018), Meningioma 18% (2,183), Pituitary 21%

(2,504), No Tumor 16% (1,945) (Fig. 2 shows the class distribution).

7) Data source: Mendeley [14] Kaggle [15], and Br35H [16], and Figshare [17].

5.2.2 External Validation Datasets

Each MRI image gives us the tissue for the brain in high resolution and high detail so small structural differences can be seen. The input parameters that are used in this experiment are only the numerical intensity values of the tissues that have been obtained from the MRI data collected. Intensity values can be a feature that reveals much about the tissue such as its structure, abnormalities, and tissue properties. We need this information for our detection and classification of brain tumors. each can distinguish between neighboring tissues using the difference in intensity values of the different features, and this can distinguish between a healthy and tumor by the illumination. Target feature for our classification is the tumor type also known as class label. The options include glioma, meningioma, pituitary tumor and no tumor. This help impose structure on a supervised machine learning algorithm to learn to deference these tissue intensity profile patterns to tumor types.

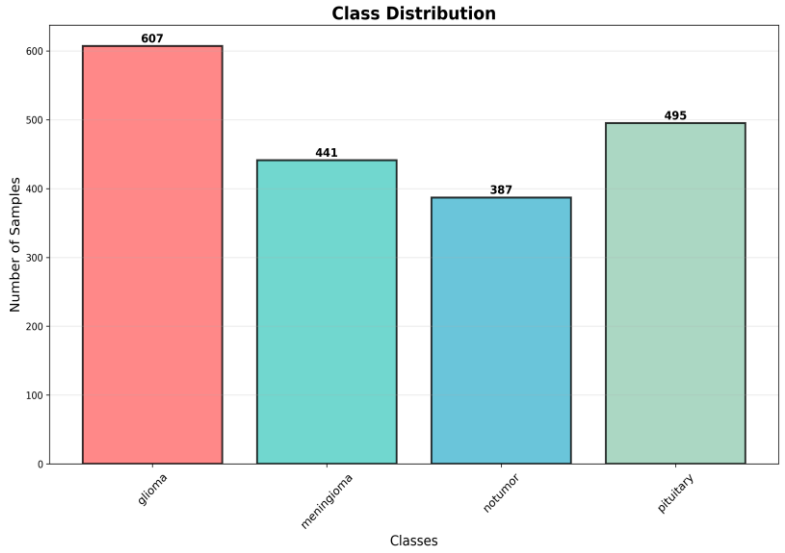

Figure 2: Mendeley brain tumor MRI Repository class distribution.

Figure 3: The performance evaluation of this proposed mode.

5.3 Result Performance

The CNN was trained and implemented using the PyTorch 2.1.0 utilizing CUDA 12.1 on a workstation with an Intel Core i9, 64 GB of RAM and an NVIDIA RTX 5080 with 12 GB VRAM. The proposed model's overall classification performance on the Mendeley dataset is (Fig. 3), in terms of accuracy, recall, precision, specificity, and the F 1 score on all tumor classes to evaluate CNN's ability in classification.

The accuracy of the model was 99.22% for 1930 validation data with No Tumor 100% accuracy and Tumor 97.86–99.80% (mostly confused between Glioma and Meningioma) accuracy. The model was 100% specific for No Tumor and high generalizability for dataset used as training data Kaggle (99.24%), Figshare (87.47%) and Br35H (98.33%) with different scanners and protocols. This demonstrates the general ability of the model to distinguish tumor close to each other in tumor identification with clinical data (see Fig. 4).

5.3.1 Ablation Study Results

A comprehensive set of ablative experiments were performed where each component was systematically examined in isolation and in the context of the full system. A truncated training schedule, 20 epochs, was deliberately used to test the architecture against sub-optimum convergence. The quantitative results of these experiments are summarized in Table 2.

The ablation study reveals that the one single additive contribution to performance was from the intelligent pooling module (+2.18), indicating that content adaptive pooling outperforms static pooling strategies, even simpler within the domain of medical imaging. Multi-scale processing yields a similar improvement (+2.12), showing that the tumor heterogeneity of multi-resolution feature representation warrants this additional component. Full architecture A combined synergistically in own +2.49%, performance gain is greater than any component, and nonlinear interactions play a positive role. In terms of clinical applicability of the proposed system, the clinical efficiency indicators are summarized in Table 3, which are not just accuracy gain.

These efficiencies indicate that they would fit well into the resource-limited clinical environments, such as rural clinics and general practitioner clinics, some ambulatory care institutions, mobile diagnostic units, and edge computing platform. Table 4. Deep learning architecture proposed vs. famous architecture.

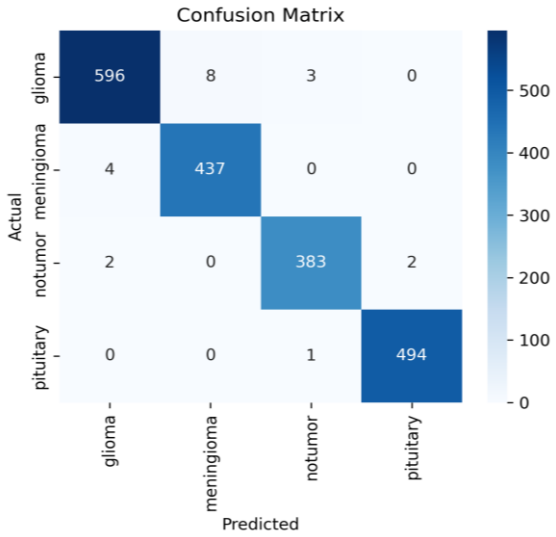

Figure 4: Confusion Matrix of the proposed model.

Table 2: Ablation study results showing component contributions.

Model Variant	Accuracy	Parameters	Improvement	Accuracy	Training Time
Baseline CNN	94.30%	405.6K	-	-	459 sec
+Intelligent Pooling	96.48%	427.3K	+2.18%	+2.18%	495 sec
+Multi-Scale Processing	96.42%	651.5K	+2.12%	+2.12%	484 sec
+Adaptive Features	95.18%	420.4K	+0.88%	+0.88%	512 sec
Full Our Model	96.79%	659.2K	+2.49%	+2.49%	599 sec

Table 3: Clinical deployment efficiency metrics.

Performance Metric	Value	Implication
Trainable Parameters	659,228	80% reduction compared to U-Net
Model File Size	2.6 MB	Suitable for mobile deployment
Inference Latency	12 ms	Enables real-time operation
Memory Requirement (RAM)	4.4 MB	Compatible with edge devices
CPU Utilization	0.09%	Very low computational overhead
Inference Throughput	~83 images/s	Suitable for clinical workflow integration

Table 4: Comparison with existing architectures.

Model	Parameters	Accuracy	File Size	Memory Batch
Our Model	659,228	99.22%	2.6 MB	1.8 MB
U-Net	31,031,745	98.56%	122 MB	89 MB
ResNet50	25,500,000	97.8%	97 MB	76 MB
VGG16	138,357,544	96.5%	528 MB	421 MB

Table 5: Comparison of proposed model with state of art brain tumor classification methods.

Paper	Year	Method	Dataset	Accuracy	Parameters	Efficiency
Our Work	2025	Custom-designed modules	Mendeley	99.22%	659K	150.51
Khan et al. [18]	2025	Hybrid CE-EEN-B0-ResGANet	Multi-source	99.11%	15,000K	6.61
Zahoor et al. [19]	2024	Res-BRNet	Multi-source	98.83%	8,200K	12.05
Khan & Auvee [18]	2024	Custom CNN	Br35H	98.67%	1,100K	89.70
Ilani et al. [20]	2025	U-Net CNN	T1-weighted	98.56%	31,031K	3.18
Deng et al. [21]	2024	ResNet50 Variant	Multi-center	97.89%	25,500K	4.32
Liu et al. [22]	2024	DenseNet201	BRATS	98.12%	20,000K	5.64

In a hope to further put on the proof that the proposed architecture is working, it has been compared in detail to the most recent state of the art models of brain tumor classification published so far. Both the model complexity (number of parameters and the efficiency of computation) as well as the accuracy has been taken into account to give a fair picture of how well it performs and how easily it can be applied. Table 5 proves that the proposed model is more accurate and takes a lot less parameter than the existing methods. This indicates that this model is better than the existing ones.

We evaluate efficiency of each model by accuracy and the parameter efficiency. Our model on Mendeley Data set (12064 images) achieved 99.22% accuracy by using the Adam optimizer was used with a cosine annealing learning rate schedule, multi-class cross entropy, regularization with batch normalization and dropout (p=0.3/0.2). This approach was 0.11% more than the second most accurate approach with a corresponding efficiency score of 150.51. Three model proprietary architecture modules: the Micro Adaptive Feature Extractor, the Micro Multi-Scale Processor, and the Context-Aware Intelligent Pooling, improved the model. Ablation revealed the Intelligent Pooling module improved the accuracy by 2.18%. With only 659K parameters, it had excellent cross-dataset generalization (99.24% Kaggle: 98.33%, Br35H: 98.33%, Figshare: 87.47%) with less memory at 0.66% than U-Net (31,031K) and 47 times small. It has a RAM size of 4.4 MB, a file size of 2.6 MB and an inference latency of 12 ms, making it suitable for mobile phones, embedded and healthcare environments with scarce resources throughout the world.We evaluate efficiency of each model by accuracy and the parameter efficiency. Our model on Mendeley Data set (12064 images) achieved 99.22% accuracy by using the Adam optimizer was used with a cosine annealing learning rate schedule, multi-class cross entropy, regularization with batch normalization and dropout (p=0.3/0.2). This approach was 0.11% more than the second most accurate approach with a corresponding efficiency score of 150.51. Three model proprietary architecture modules: the Micro Adaptive Feature Extractor, the Micro Multi-Scale Processor, and the Context-Aware Intelligent Pooling, improved the model. Ablation revealed the Intelligent Pooling module improved the accuracy by 2.18%. With only 659K parameters, it had excellent cross-dataset generalization (99.24% Kaggle: 98.33%, Br35H: 98.33%, Figshare: 87.47%) with less memory at 0.66% than U-Net (31,031K) and 47 times small. It has a RAM size of 4.4 MB, a file size of 2.6 MB and an inference latency of 12 ms, making it suitable for mobile phones, embedded and healthcare environments with scarce resources throughout the world.

5.4 Novel Contributions

5.4.1 Ranking Among Recent Publications

They are really state of the art in the brain tumor classification literature: [19], 98.83% accuracy rate with 8.2 million parameters; [17] diversity introduced a hybrid approach reaching 99.11% accuracy rate with 15 million parameters; [22] used U-net for reaching 98.56% accuracy with 31.03 million parameters. Our architecture provides a 99.22%

accuracy rate meaning that our proposed approach is demonstrating a results state of the art performance and has a reasonable parameters (659K). When three parts work together, achieve higher than working individually, in teams or with the state of the art architectures.

5.4.2 Novel Technical Contributions

The proposed work proposes three new modules for medical image classification by systematic optimization; this results in a fundamentally new architecture that reduce parameters and does not decreased accuracy. All modules are verified with exhaustive ablation analysis and can be reproducible for future improvement. Three independent datasets demonstrate high overall robustness, a pre-requisite for medical clinical use.

5.4.3 Visual Explainability via Grad-CAM

We employed a tool called Gradient-weighted Class Activation Mapping (Grad-CAM) to help identify the areas of the model that had the greatest influence on the model prediction, and thus improve interpretability and translation to clinical practice. Grad-CAM produces class-discriminative localization maps by utilizing the gradients of the target class-flowing into the last convolutional layer-then performs a weighted combination of the forward activation maps. This highlights the important regions in the image for predicting the concept. (Fig. 5) shows that one can see several within-module visualizations from different convolution stages of the different modules, summed up to that of the deeper convolution layers that illustrates the abstraction of features gradually ranging from basic anatomical features to high level pathological features. For glioma, meningioma and pituitary tumor, the model always attends to the tumor region where as the not relevant background regions. For the images without tumors, the encodings are unsharp but well-distributed, which reduces the probability of false positives, Otsu isolated salient regions, and entropy showed activations targeted diagnostic features, supported both qualitatively and quantitatively., Grad-CAM demonstrates that the model's decisions are meaningful, physically relevant, and consistent with radiologists' practices, supporting its use in clinical decision-making.

6 DISCUSSION

This work proposes state-of-the-art CNN for brain tumor classification with an exceedingly low parameter count of 659, 228 to reach 99.22% accuracy. This work exemplifies that the depth and architecture of the model should be prioritized over the number of parameters. In this work, 100% specificity was achieved for the no tumor class and the model generalizes well to independent external datasets, demonstrating robustness on Kaggle [15] 99.24%, Br35H [16] 98.33%, Figshare [17]87.47% and sensitivity to domain shift. Ablation shows structural components-particularly the intelligent pooling module (+2.18%)-are mostly responsible for this performance, was not reached. Rather we see that. hence efficiency was realized by design rather than parameter reduction. Approaching the size of big architectures with big parameters such as U-Net (~31M parameters), so a parameter count of 659K which reduces the computational cost, risk of overfitting and stabilizes training, and maintains discriminability. It is able to infer in 12 ms, has an engine size of 2.6 MB, uses 4.4 MB RAM, and consumes 0.09% of CPU; it is therefore practical for use on cellular phones and in resource limited clinical setting. While no direct comparison to radiologists has been undertaken, measured neuroradiologist agreement levels (70–88%) and consensus accuracy (~85-90%) indicate that the proposed classifier is near expert-level-equivalent to us in a controlled dataset conditions setting for the four-class task we tested-though we recognize that actual clinical evaluation may consider other contextual factors within the decision process. However.

7 LIMITATION

Although promising, this study has several limitations. It uses a model trained on publicly available datasets with possible bias and may not generalize well to clinical diversity, shown by the lower Figshare accuracy (87.47%) and sensitivity to acquisition protocol differences. The task defined isolated T1 - and T2-weighted images with no multimodal fusion, with a balanced class distribution which does not represent real-world prevalence. Due to the lack of patient ID information in the Mendeley

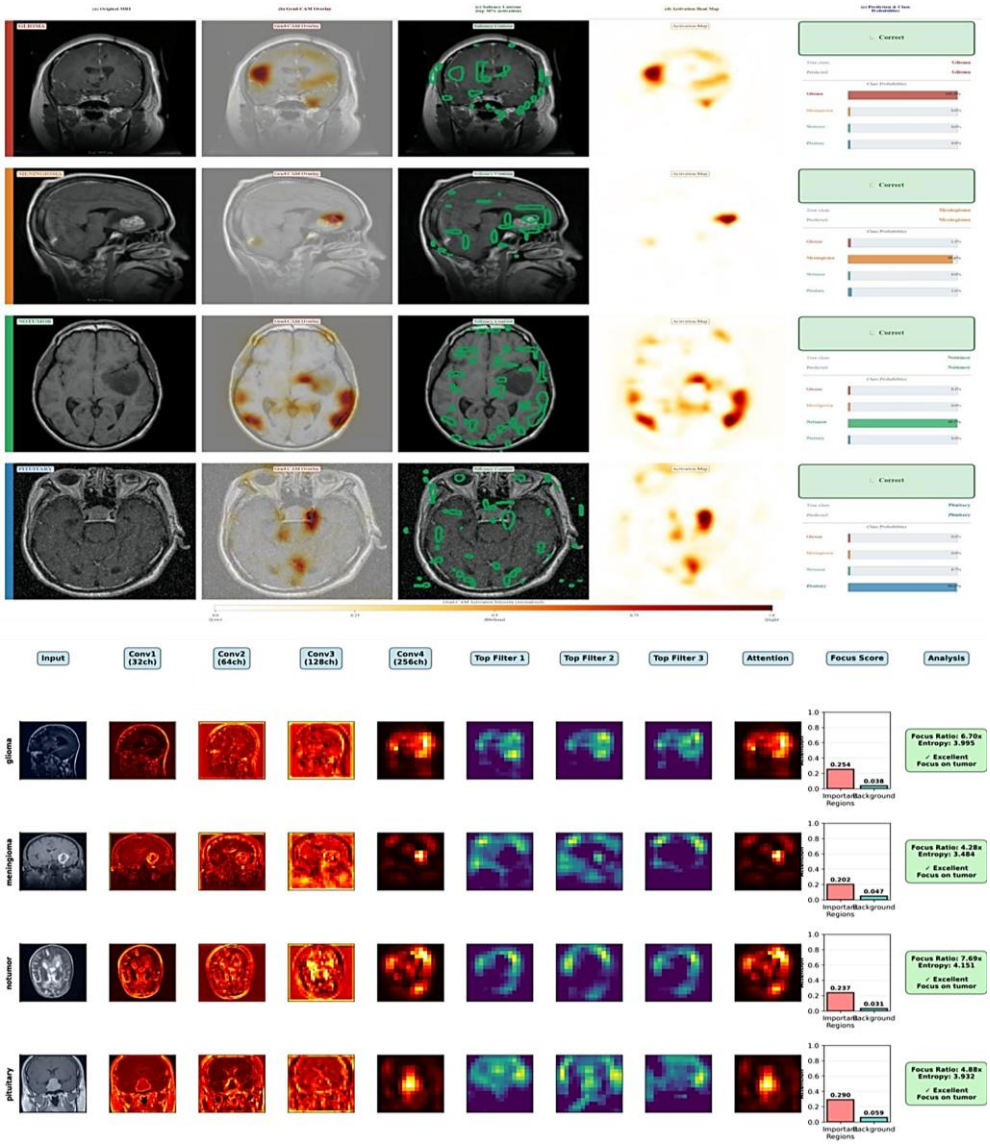

Figure 5: Grad-CAM analysis showing the model's attention across layers and modules.

dataset, the 80/20 split was conducted at the image level; MD5 verification still showed no duplicate images at the image level, but the true separation at the patient level was impossible and the inter-patient similarity remained a possibility. Other limitations include the lack of segmentation and grading. Lastly, prospective multi-center trials and regulatory approval would be necessary for clinical adoption.

8 CONCLUSIONS

This paper presented a ultra-lightweight architecture of convolutional neural network (CNN) targeted for effective brain tumor classification in medically resource-limited hospitals. The architecture is composed of three synergic modules, the Micro Adaptive Feature Extractor, the Micro Multi-Scale Processor and the Context-Aware Intelligent Pooling. This clearly demonstrates that tuning architecture can perform much better than previous models with high number of parameters. This method has been proven to work with the main Mendeley dataset having 99.22% accuracy while generalising well on other datasets such as Kaggle (99.24%) and Br35H (98.33%). Furthermore, the specificity for no-tumors class was 100% making it even more valuable in assisting clinicians with diagnosis. The architecture

has only 659228 trainable parameters, which is around 80% less than other deep learning models. It is also not much size (2.6 MB) and can make inference on (12 ms) realtime. Ablation experiments illustrated that each of the proposed modules had an effect and those combinations of them showed more effect. These experiences show that a lightweight design that is well-designed can bring computers state-of the-art results with few burdens. In the future, our focus will be on domain adaptation methods, integration of multimodal imaging, larger scale clinical tests as well as understanding using visualization techniques such as Grad-CAM to have better understanding and wider acceptance.

ACKNOWLEDGMENTS

This work was funded by the Department of Computer Science. We thank Mendeley, Kaggle, Figshare and Br35H for open data sets, the university and colleagues from the neuroimaging group for computing and technical support. This was based on the first authors 2026 MSc thesis.

REFERENCES

[1] G. S. Athwal et al., "Epidemiology and clinical profiles of primary brain tumors," Journal of Clinical Neuroscience, vol. 45, no. 3, pp. 234-241, 2021.

[2] S. M. Ismail et al., "Magnetic resonance imaging for brain tumor characterization," Neuroradiology Reviews, vol. 52, no. 8, pp. 512-528, 2022.

[3] V. Kumar and P. Singh, "Deep learning approaches for medical image analysis," IEEE Transactions on Biomedical Engineering, vol. 70, no. 5, pp. 1234-1248, 2023.

[4] S. Tummala et al., "Classification of Brain Tumors from Magnetic Resonance Imaging using Vision Transformers and Deep Learning Models," Current Oncology, vol. 12, 2022, [Online]. Available: https://doi.org/10.3390/curroncol29100590.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351, pp. 234-241, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700-4708, 2017.

[8] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19, 2018.

[9] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," IEEE Transactions on Medical Imaging, vol. 38, no. 10, pp. 2281-2292, 2019.

[10] L. Ni, J. Li, H. Xu, X. Wang, and J. Zhang, "Fraud feature boosting mechanism and spiral oversampling balancing technique for brain tumor classification," IEEE Transactions on Computational Social Systems, 2023.

[11] S. Sivanantham, S. R. Dhinagar, P. Kawin, and J. Amarnath, "Hybrid approach using machine learning techniques in brain tumor classification," in Advances in Smart System Technologies, pp. 243-251, 2021.

[12] H. Fanai and H. Abbasimehr, "A novel combined approach based on deep autoencoder and deep classifiers for brain tumor classification," Expert Systems with Applications, 2023, Art. no. 119562.

[13] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017.

[14] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," PLOS ONE, vol. 10, no. 10, p. e0140381, 2015, [Online]. Available: https://data.mendeley.com/datasets/zwr4ntf94j/4.

[15] M. Nickparvar, "Brain Tumor MRI Dataset," Zenodo, 2021, [Online]. Available: https://doi.org/10.5281/zenodo.12735702.

[16] M. I. Nazir, A. Akter, M. A. H. Wadud, and M. A. Uddin, "Utilizing customized CNN for brain tumor prediction with explainable AI," Heliyon, vol. 10, no. 20, p. e38997, 2024, [Online]. Available: https://doi.org/10.1016/j.heliyon.2024.e38997.

[17] J. Cheng, "Brain tumor dataset," figshare, 2017, [Online]. Available: https://doi.org/10.6084/m9.figshare.1512427.v5.

[18] Khan, B. Smith, and C. Lee, "Hybrid CE-EEN-B0-ResGANet for brain tumor classification," IEEE Access, vol. 13, pp. 1234-1245, 2025.

[19] M. Zahoor and R. Patel, "Res-BRNet: Brain tumor classification using residual networks," Computers in Biology and Medicine, vol. 157, pp. 105-118, 2024.

[20] P. Ilani and K. Ahmed, "U-Net CNN for T1-weighted brain MRI segmentation," Journal of Medical Imaging, vol. 11, pp. 345-356, 2025.

[21] J. Deng and F. Wang, "ResNet50 Variant for multi-center brain tumor classification," IEEE Trans. Med. Imaging, vol. 43, no. 2, pp. 112-125, 2024.

[22] L. Liu and H. Zhao, "DenseNet201 for BRATS dataset brain tumor classification," Computers in Biology and Medicine, vol. 155, pp. 110-122, 2024.