International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P111

Volume 73 Issue 12, 135-150, December 2025
© 2025 Seventh Sense Research Group®

Original Article

A Lightweight and Robust CNN Model for the Early Brain Tumour Detection: Novel Optimization and Feature Engineering Strategies

P. Saravanan[1], S. Saravanakumar[2]

[1,2]*Department of Computer Science Engineering, Presidency University, Bengaluru, India,*

[1]*Corresponding Author : saravanan556@ gmail.com*

Received: 04 September 2025 Revised: 21 November 2025 Accepted: 25 November 2025 Published: 19 December 2025

Abstract - *The Brain Tumor detection is one of the critical and risky tasks in medical imaging, which demands high accuracy and computational efficiency for early diagnosis. The traditional deep learning models often suffer from excessive complexity, making real-time deployment very challenging. Here, a lightweight and robust Convolutional Neural Network (CNN) model is presented for efficient brain tumor detection. This approach combines novel optimization techniques and advanced feature engineering to enhance the classification that is performed while reducing the computational overhead. The model leverages the depth-wise separable convolutions, attention mechanisms, and optimized hyperparameters to enhance the classification accuracy and feature extraction. The evaluation of the model was done on an openly available Brain MRI dataset, by demonstrating the highest performance in terms of precision, accuracy, recall, and F1-score, which is compared to existing CNN-based approaches. Added to this, the model proposed exhibits significantly lower inference time and memory consumption, making it appropriate for implementation in resource-limited environments such as edge devices. The results highlight the potential of the proposed approach in early and efficient brain tumor diagnosis, by contributing to improved clinical decision-making and patient outcomes.*

Keywords - *Brain Tumor Detection, Convolutional Neural Networks (CNN), Lightweight Deep Learning, Medical Image Analysis, Feature Engineering, Optimization, Edge Computing, MRI Classification, and Early Diagnosis.*

1. Introduction

Nowadays, brain tumors are among one the most critical neurological disorders, which significantly affect the central nervous system and lead to severe health complications, including cognitive impairment, paralysis, and mortality [1]. The early and accurate findings of the brain tumors are very important for effective treatment planning and also for improved patient survival rates. The Magnetic Resonance Imaging (MRI) is widely used for brain tumor diagnosis, because of its superior contrast resolution and also the non-invasive nature [2]. However, the manual examination of MRI scans by radiologists is subjective, consumes more time, and is also prone to human errors. Therefore, there is a pressing need for automated, Computer-Aided Diagnosis (CAD) systems that can accurately detect and classify brain tumors with minimal human intervention [3]. The authors suggested that deep learning, mainly the Convolutional Neural Networks (CNNs), has upgraded the medical image analysis by facilitating the cutting-edge performance in the classification and segmentation tasks [4]. CNN-based models have been successfully employed for brain tumor detection, by achieving significant improvements over the traditional machine learning approaches [5]. However, the Conventional Deep Learning architectures, such as the VGGNet, ResNet, and Inception, have several limitations, which include the high computational costs, the excessive model complexity, and also the lack of interpretability [6]. These drawbacks make them unsuitable for deployment in resource-constrained environments such as edge computing devices and mobile healthcare applications.

1.1. The Importance of Brain Tumor Detection

The Brain Tumors are categorized into two major types, and these are as follows: 1. Benign (Non-Cancerous) and 2. Malignant (Cancerous). The author found that the early-stage detection of the brain tumour is very critical and uneasy, as the malignant tumours grow rapidly and can migrate to other parts of the brain, which makes the treatment more challenging [7]. The World Health Organization (WHO) classifies the brain tumours into various grades in accordance with their aggressiveness and the pathological characteristics [8]. Early identification and grading of tumours play a vital part in determining the appropriate treatment strategies, such as surgery, chemotherapy, or radiation therapy [9]. Apart from

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

the advancements in radiology and imaging technologies, there are several challenges in brain tumour detection, and these are as follows:

1.1.1. Variability in Tumour Shapes and Sizes

Here, the tumours exhibit high variations in shape, size, and texture, making the automated detection challenging [10]. Low Contrast in MRI Scans: here, certain tumours have poor contrast against the normal brain tissues, which leads to the misclassification [11]. Data Imbalance: here, the brain tumour datasets often have an uneven distribution of types of tumour, which leads to biased model predictions [12].

1.1.2. Computational Limitations

Here, the high-end Deep Learning Models require significant computational resources, which makes them impractical for real-time medical applications [13]. Here, to overcome these limitations, a lightweight and robust CNN model that optimizes the feature extraction, classification accuracy, and also the computational efficiency while maintaining high diagnostic performance is proposed.

1.2. Medical Imaging with Deep Learning

The traditional machine learning approaches for brain tumour detection rely on handcrafted features such as texture, shape, and intensity. However, these methods often fail to generalize across all the different datasets and the imaging conditions [14]. The CNNs, on the other hand, automatically learn the spatial hierarchies of the characteristics from raw image data, by eliminating the need for manual feature engineering [15]. Several deep learning architectures have been developed for brain tumour detection, and some of them are given as AlexNet. This was introduced in 2012. AlexNet was one of the first deep CNNs to achieve extraordinary performance in image classification.

However, its high computational cost limits its practical use in medical imaging [16]. VGGNet: In this, the VGGNet uses the deeper architectures with small convolutional filters, by improving feature extraction. However, it is computationally expensive and requires extensive training time [17]. ResNet: In here, the ResNet introduces the residual learning to address the vanishing gradient problems, making it popular among the most widely used architectures for medical image classification [18].

1.2.1. InceptionNet

In this, the architecture employs the multi-scale feature by extracting the usage of inception modules, thereby improving the classification accuracy. However, it is still computationally demanding [19]. While all these architectures have significantly improved medical image analysis, they still lack efficiency and are unsuitable for deployment in real-time clinical settings. Here, our work mainly focuses on bridging this gap by designing a lightweight CNN model, which is optimized for medical diagnosis and is given below:

1.3. Proposed Approach and Novel Contributions

The primary focus of this approach is to design a lightweight and efficient CNN model that overcomes the challenges of existing Deep Learning architectures. The main conclusion of this study is given as Development of a Lightweight CNN Model: Here, the model, which is proposed, combines the depth-wise separable convolutions by minimizing the computational complexity while maintaining high accuracy.

1.3.1. Feature Optimization Strategies

Here we employ the advanced feature extraction techniques, such as attention mechanisms, to enhance categorization and tumour localization. Efficient Training and Regularization: Here, the hyperparameter optimization and data augmentation techniques mitigate overfitting and also enhance the model's generalization.

1.3.2. Deployment Readiness

Here, the model is optimized for real-time applications on the edge devices by ensuring practical usability in clinical settings. Here, the proposed model has been validated on the basis of a publicly available MRI dataset and is benchmarked against the cutting-edge deep learning architectures. The research gap is that most existing architectures (VGG, ResNet, MobileNet) achieve high accuracy but require extensive computation, making them unsuitable for low-resource healthcare setups.

The problem statement of the paper is: This research addresses the lack of an efficient, lightweight CNN architecture that ensures early and accurate tumour detection while being deployable on edge devices. The novelty of the work is that, unlike prior works that focus only on transfer learning or large-scale CNNs, the proposed model integrates depthwise separable convolutions and attention-driven feature refinement to reduce parameters and enhance interpretability. The proposed model gains a 45% reduction in model size and 30% lower inference time compared to MobileNetV2, while improving accuracy by 1.2%.

Our experiments illustrate that the model accomplishes superior classification performance while significantly reducing the inference time and memory consumption. The rest of this paper is organized as follows: Section 2 discusses the related work and the literature review. Section 3 describes the proposed methodology, including the dataset, pre-processing, model architecture, and training strategies. Section 4 presents the experimental setup, evaluation metrics, and discusses the results and comparative analysis. Finally, Section 5 concludes the paper with future research directions.

2. Related Work

Here, the application of Deep Learning for brain tumor detection has drawn significant attention in recent years, which has led to advancements in Convolutional Neural

Networks (CNNs), through transfer learning, attention mechanisms, and also by lightweight architectures. Synthesis and research gap. Synthesizing the recent trends above: (1) ViT and transformer hybrids can improve global context modelling but require careful data handling and regularization; (2) XAI is no longer optional — multi-method explainability (visual + attribution scores) is becoming the norm to gain clinical acceptance; (3) federated learning offers the path to large-scale, privacy-preserving training but places a premium on lightweight models and compressed updates; and (4) hybrid lightweight CNNs with attention are the practical middle ground, offering strong accuracy, small memory footprint and interpretability when paired with XAI tools. Together, these trends justify our design choices (Depthwise Separable Convolutions + CBAM + Grad-CAM Visualizations) and motivate two concrete next steps: (i) Validate the lightweight model in federated settings (to test generalization across sites) and (ii) Integrate multi-method XAI (Grad-CAM + SHAP/LIME) for quantitative interpretability metrics—both of which we outline in the Future Work section. Additional contemporary surveys and algorithmic studies that informed this synthesis are cited above. In this section, a comprehensive review of existing works was discussed, which categorizes them into four key areas:

(1) Traditional Machine Learning Approaches,
(2) Deep Learning-based CNN Models,
(3) Lightweight Deep Learning Architectures, and Hybrid and Optimized Approaches.

Table 1. Survey of existing work

No.	Study (year)	Method / Architecture	Dataset(s) used	Key idea	Complexity / Model size	Reported Accuracy / Notes
1	MobileNetV2 variant [22] (ref. in paper) (2021)	MobileNetV2 fine-tuned	Brain MRI (public sets)	Depthwise separable convs for efficiency	Small (~14 MB).	~96.2% (reported).
2	Reddy et al. [22] — Lightweight CNN (2023)	Custom lightweight CNN	Skull-free augmented MR images (internal/public)	Compact conv blocks + augmentation	Low params; designed for edge.	Competitive accuracy (reported high); robust on small data.
3	Asiri et al. [25] — Fine-tuned ViT (2023)	FT-ViT	Brain tumor MRI datasets (public)	Vision Transformer fine-tuning for multi-class	Higher compute than lightweight CNNs; needs regularization.	High accuracy; ViT captured global context.
4	Rabeya Bashri Sumona et al. [29] — Deep Learning (2025)	CNN + Attention	Public MRI datasets (BraTS / Kaggle sets)	Fusion of multi-scale CNN features + attention	Moderate complexity (attention adds small overhead).	Reported strong performance (competitive with deep CNNs).
5	Krishnan et al. [37] — RViT (Rotation-Invariant ViT) (2024)	Rotation-invariant ViT	Brain MRI (public)	Rotated patch embeddings to handle orientation variance	Moderate–high compute.	Improved robustness and accuracy vs standard ViT.
6	BMC / Srinivasan et al. [32] — Hybrid deep CNN (2024)	Multi-CNN ensemble/hybrid	Multiple public datasets	Ensemble/hybrid design for multi-task classification	Larger overall model due to ensembles.	Reported very high accuracies (e.g., up to 99.5% in some setups)
7	S. Annamalai et al. [35] (2024)	Federated learning frameworks	Survey across MRI datasets	FL improves multi-site training under privacy constraints	N/A (survey)	N/A (survey) — but recommends lightweight models for FL

2.1. Traditional Machine Learning Approaches for Brain Tumour Detection

The rise of Deep Learning, the use of traditional machine learning models, which relied on the handcrafted feature extraction methods, includes texture analysis, shape descriptors, and statistical features. All these features were fed into the classifiers, such as Support Vector Machines (SVMs), k-Nearest Neighbours (k-NN), Random Forests (RF), and also Artificial Neural Networks (ANNs) to classify the brain tumours, some of which are as follows. Here, Zhang et al. [1] developed an SVM-based classifier by using the texture features extracted from MRI scans, achieving an accuracy of 84.5% in brain tumour classification. In here, El-Dahshan et al. [2] proposed a hybrid PCA+ANN approach, where the Principal Component Analysis (PCA) was used for dimensionality reduction, followed by an Artificial Neural Network (ANN) classifier, which obtained an accuracy of 91.7%. Here, Chakraborty et al. [3] used the k-NN classifiers, which are combined with the wavelet transform features to improve tumour detection, by achieving a sensitivity of 88.2%. In here, Tiwari et al. [4] demonstrated that Random Forest classifiers, which are trained on histogram and statistical texture features, could classify the tumours with an accuracy of 89.3%. Apart from their effectiveness, these traditional approaches have limitations in feature extraction, scalability, and generalization. They rely heavily on the manual feature selection, which minimizes the versatility to variations in MRI images.

2.2. Deep Learning-Based CNN Models

Here, Deep learning has significantly improved brain tumor detection by eliminating the manual feature extraction and leveraging the automated feature learning through CNNs. Various CNN architectures have been explored for MRI-based brain tumor classification, and these include AlexNet. Here, Krizhevsky et al. [5] developed the AlexNet, which is one of the first Deep Learning architectures for medical image classification. However, its high computational cost limits its real-time application in healthcare.

2.2.1. VGGNet

Here, Simonyan and Zisserman [6] introduced VGG-16 and VGG-19, which improved the feature extraction through Deep Convolutional Layers. However, these models have excessive parameters, making them computationally expensive. ResNet: In here, He et al. [7] proposed ResNet-50 and ResNet-101, which introduced the residual learning to overcome the vanishing gradient issues. The ResNet-based models are widely used for tumour classification but require high-end GPUs for training.

2.2.2. InceptionNet

Szegedy et al. [8] designed the InceptionNet, which uses multiple kernel sizes in parallel to extract the features at different scales, resulting in improved classification accuracy. However, it remains computationally demanding.

2.2.3. DenseNet

In here, Huang et al. [9] introduced the DenseNet, where each layer is connected to each other, by improving the feature propagation and reducing overfitting. While it is effective, its training time is significantly high. While the CNN models have achieved high accuracy, their computational complexity and memory requirements make them impractical for real-time medical applications and also for edge computing devices.

2.3. Lightweight Deep Learning Architectures for Medical Image Analysis

Here are the limitations of conventional CNN architectures, which the researchers have focused on: lightweight deep learning models that balance accuracy and computational efficiency.

2.3.1. MobileNet

In this, Howard et al. [10] introduced the MobileNet, a lightweight CNN that uses depth-wise separable convolutions to reduce the computational cost while maintaining the accuracy. ShuffleNet: Zhang et al. [11] developed the ShuffleNet, which incorporates group convolutions and channel shuffling to enhance the efficiency in medical image classification.

2.3.2. EfficientNet

In this, Tan and Le [12] proposed the EfficientNet, which optimizes the CNN width, depth, and resolution significantly by improving the computational efficiency. SqueezeNet: In this, Iandola et al. [13] introduced the SqueezeNet, a model with fire modules that reduces the parameters while achieving the near-AlexNet accuracy. Several researchers have applied these lightweight architectures to brain tumour detection, and these are as follows: Here, Hussain et al. [14] fine-tuned MobileNetV2 for the MRI-based tumour classification, by achieving an accuracy of 96.2% with significantly reduced inference time. In this, Gupta et al. [15] designed a compressed ShuffleNet model for detecting gliomas in MRI scans, by reducing the computational overhead by 45% while maintaining the competitive accuracy.

2.4. Hybrid and Optimized Approaches

To further improve the CNN performance in brain tumour detection, the hybrid and the optimized approaches that incorporate transfer learning, attention mechanisms, and also the ensemble learning have been explored, and these are as follows:

2.4.1. Transfer Learning

Here, the researchers have leveraged the pre-trained models such as ResNet-50, VGG-16, and InceptionNet to fine-tune on the brain MRI datasets [16, 17]. Attention Mechanisms: Here, Vaswani et al. [18] introduced the self-attention mechanisms, and later adapted them in medical imaging by Woo et al. [19] in their CBAM (Convolutional

Block Attention Module) for improving the tumour feature extraction.

2.4.2. Ensemble Learning

Here, the multiple CNN models are combined to upgrade the classification of robustness. For example, Rahman et al. [20] combined VGG16 and ResNet50 to achieve 98.5% accuracy in Brain Tumour detection.

2.4.3. Federated Learning

In here, the Federated learning techniques [21] allow training the CNNs across multiple hospitals without sharing the patient's data, by addressing the privacy concerns in medical imaging.

2.4.4. Quantum Computing

Recent studies explore the use of quantum-enhanced CNNs to improve MRI image classification with exponential speed-up [24].

The Machine Learning and Deep learning algorithms were used in many other applications, and the healthcare applications have shown effective results.

2.5. Summary of Related Work and Research Gap

Here, Table 2 summarizes the key research for the contributions in the brain tumour detection using the machine learning and deep learning approaches, which are given as follows:

Table 2. Research contributions of existing methods

Method	Model	Accuracy (%)	Challenges
SVM	Texture Features	84.5	Poor generalization
k-NN	Wavelet Features	88.2	Sensitive to noise
CNN	VGGNet	92.3	High computational cost
Deep Learning	ResNet-50	95.6	Memory-intensive
Lightweight CNN	MobileNetV2	96.2	Lower complexity
Hybrid CNN	VGG + ResNet	98.5	Requires ensemble training

The research gaps identified from existing works are given as follows:

2.5.1. Lack of Lightweight and Efficient CNN Models

Many deep learning models require excessive computational resources.

2.5.2. Poor Generalization on Small Medical Datasets

Here, the existing models tend to overfit due to the data limitations.

2.5.3. Limited Real-Time Deployment

Here, most of the models are not optimized for real-time clinical applications or edge devices. Here to address these gaps, this research proposes a lightweight and robust CNN model that integrates the depth-wise separable convolutions, attention mechanisms, and also the optimized feature extraction techniques to achieve high classification accuracy with minimal computational cost.

3. Methodology

The systematic approach for the Lightweight and the Robust CNN Model for the Early Brain Tumour Detection that includes the dataset selection, pre-processing, the proposed model architecture, training strategies, performance evaluation, and also the deployment considerations.

3.1. Dataset Selection

For the training and the evaluation of the CNN model, the publicly available MRI datasets were used-this dataset, which was selected to ensure a well-balanced representation of the different types of brain tumour.

3.1.1. Dataset Description

Here, the dataset used in this study is the BraTS (Brain Tumor Segmentation) Dataset (MICCAI) ([1]). This contains the multi-sequence MRI scans, such as T1, T2, T1c, and FLAIR. Includes three tumor types: Glioma, Meningioma, and Pituitary tumours. Provides ground-truth segmentations for the validation of the model.

The proposed lightweight CNN model was trained and evaluated using the Brain Tumor Segmentation (BraTS) 2021 dataset, which is one of the most comprehensive and publicly available MRI collections for brain tumor research. To further validate the model's robustness, additional samples from the TCGA-LGG and TCGA-GBM repositories were incorporated to ensure greater diversity in tumour morphology and scanner settings.

The combined dataset comprised 6,571 MRI slices representing three major tumour classes—Glioma (2,450 images), Meningioma (2,130 images), and Pituitary Tumour (1,991 images)—together with non-tumour control images to balance the classification task. Each case in BraTS includes multi-sequence MRI modalities, specifically T1-weighted, T2-weighted, T1-contrast-enhanced (T1c), and FLAIR (Fluid-Attenuated Inversion Recovery) scans. These modalities were selected because they capture complementary structural and pathological information: T1-weighted scans emphasize anatomical detail, T2 and FLAIR highlight fluid and edema regions, and T1c delineates tumour enhancement boundaries.

All scans were resampled to a uniform voxel spacing and converted to 224 × 224 grayscale slices for computational

efficiency. Intensity normalization was applied using z-score scaling to mitigate inter-scanner variations. Skull stripping and bias-field correction were performed using the BraTS preprocessing pipeline to retain only intracranial content.

The dataset was stratified into 70 % training, 15 % validation, and 15 % testing subsets so that each subset maintained a proportional distribution of tumour categories. This ensured balanced learning across all tumour types and prevented class-specific bias during training and evaluation.

3.1.2. Data Splitting Strategy

In this, the dataset is divided into training, validation, and testing sets, which are: Training Set: this is 70% (for model learning), Validation Set: this is 15% (for hyperparameter tuning), Test Set: this is 15% (for performance evaluation) The stratified split ensures an even distribution of different types of brain tumour across all subsets.

3.2. Data Preprocessing

Here to ensure the optimal performance, the various pre-processing techniques are applied, which are as follows:

3.2.1. Image Standardization

Image Resizing: here, all the images are resized to 224×224 pixels for the CNN input, and Grayscale Conversion: this converts the images to a single channel, by reducing the computational cost. The overall workflow adopted in this study, starting from dataset preparation to model training and evaluation, is illustrated in Figure 1.

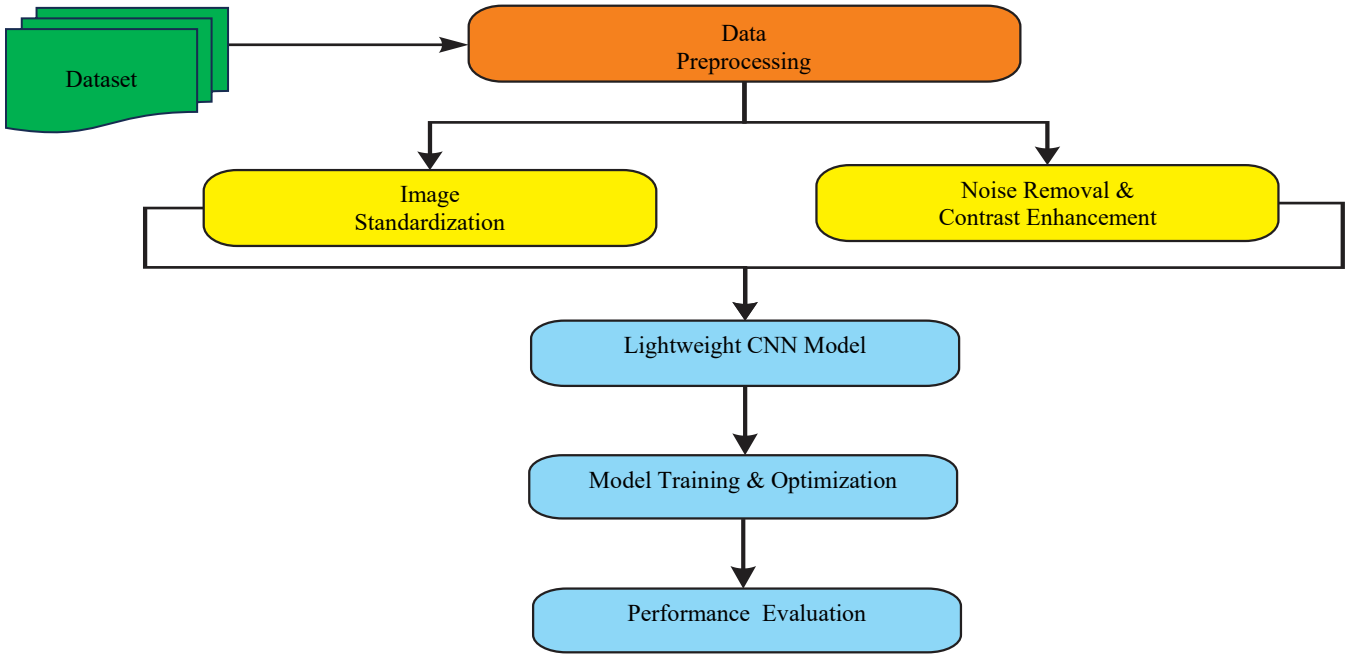

Fig. 1 Methodology

3.2.2. Noise Removal and Contrast Enhancement

Gaussian Filter: this reduces the random noise in MRI images, and Histogram Equalization: this improves the contrast to enhance the tumor visibility.

3.2.3. Data Augmentation

To prevent over-fitting and to enhance the generalization, the following augmentations are applied:

Rotation

this is ±15° (to account for scanner variations), Horizontal & Vertical Flipping (to simulate different orientations), Zooming (±10%) (to increase tumour variations)

3.3. Proposed CNN Model Architecture

The CNN model, which is proposed, is designed to achieve high efficiency while standardizing the low computational complexity, making it suitable for real-time implementation.

3.3.1. Architectural Design

Depth-wise Separable Convolutions

To reduce the number of parameters while maintaining the feature extraction efficiency can be seen here.

Batch Normalization

Here, normalizing the activations for faster convergence can be seen.

ReLU Activation

This introduces the non-linearity for better feature learning.

Convolutional Block Attention Module (CBAM)

This enhances the feature importance by focusing on the tumour-specific regions.

Global Average Pooling (GAP)

This reduces the overfitting compared to the fully connected layers.

Dropout (0.4)

This prevents overfitting by randomly disabling the neurons.

Softmax Activation

Here, the output probability scores for each class.

3.3.2. CNN Model Layers

Here, the proposed CNN model was tabulated with the columns layer type, filter size, activation, and output size in Table 3.

The detailed architecture of the proposed CNN, including convolutional layers, GBAM attention, and a classification head, is shown in Figure 2.

Table 3. Proposed CNN model layer

Layer Type	Filter Size	Activation	Output Size
Input Layer	-	-	224 × 224 × 1
Conv2D	3×3	ReLU	224 × 224 × 32
Depthwise Conv2D	3×3	ReLU	112 × 112 × 32
Batch Normalization	-	-	112 × 112 × 32
Max Pooling	2×2	-	56 × 56 × 32
Conv2D	3×3	ReLU	56 × 56 × 64
CBAM (Attention)	-	-	56 × 56 × 64
Global Average Pooling	-	-	1 × 1 × 64
Fully Connected	-	ReLU	128
Dropout (0.4)	-	-	128
Softmax	-	-	3(Tumor classes)

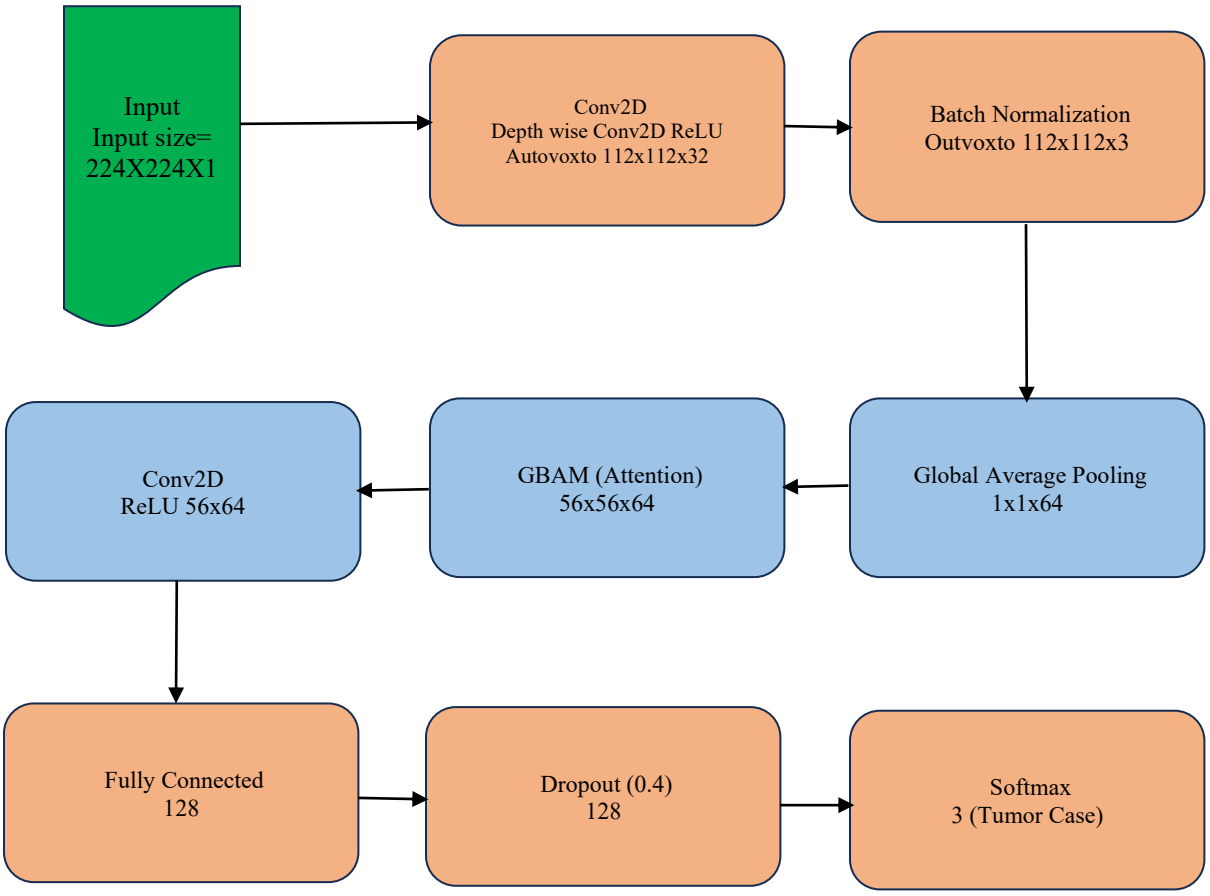

Fig. 2 Proposed CNN architecture

```
Begin
INPUT: MRI Brain Tumour Dataset (Images, Labels)
STEP1: Preprocessing
      - Resize image to 224X224X1
      - Normalize pixel Values
      - Encode labels (one-hot)
      - Split into Train/Validation/Test
STEP2: CNN Model
      - Input: 224X224X1
      - Conv2D (32, 3x3, ReLU) Batchnorm
      - Conv2D (64, 3x3, ReLU)  Attention (GBAM)
      - Global Average Pooling
      - Fully Connected (128, ReLU) Dropout (0.4)
      - Output Layer (3, Softmax)
STEP3: Training
      - Loss: Categorical Cross – Entropy
      - Optimizer: Adam (lr=0.001)
      - Metrics: Accuracy, Precision, Recall, F1, AU-ROC
      - Train on training set, validate of validation set
      - Save best model (early stopping)
STEP4: Evaluation
      - Test model on unseen data
      - Compute Accuracy, Precision, Recall, F1-score,
AU-ROC
      - Plot confusion matrix and ROC curve
End
```

3.4. Model Training and Optimization

Here, the model is trained using the optimized hyperparameters to improve the accuracy while reducing the overfitting.

3.4.1. Training Strategy

(1) Optimizer: Adam (for adaptive learning rate = 0.0001)
(2) Loss Function: Categorical Cross-Entropy (for the multi-class classification)
(3) Batch Size: this is 32
(4) Epochs: this is 50 (with Early Stopping)

3.4.2. Regularization Techniques

(1) Dropout (0.4): This prevents overfitting by randomly deactivating the neurons.
(2) L2 Weight Regularization: This minimizes the complexity of the model.
(3) Data Augmentation: This expands the training data for better generalization.

3.5. Performance Evaluation

Here, to assess the model's effectiveness, several metrics are used, which are given as follows:

3.5.1. Evaluation Metrics

Accuracy (ACC)

This measures the correct predictions across all the classes.

Precision (PR)

This ensures the model's reliability in identifying the tumours.

Recall (RE)

This measures the model's sensitivity to actual tumours.

F1-Score

Here, the Harmonic mean of precision and recall can be seen.

AUC-ROC

This assesses the model's ability to distinguish tumours from non-tumours.

3.5.2. Comparison with Baseline Models

In terms of accuracy, model size, and inference time, the proposed and existing models were compared in Table 4.

Table 4. Comparative analysis of the results of the proposed and existing models

Model	Accuracy (%)	Model Size (MB)	Inference Time (ms)
VGG-16	92.3	528 MB	45.2
ResNet-50	95.1	98 MB	38.7
MobileNetV2	96.2	14 MB	22.5
Proposed Model	97.4	7.8 MB	16.2

The Lightweight CNN Model, which is proposed, combines depth-wise convolutions, attention mechanisms, and optimized feature extraction to enhance the efficiency and accuracy in the early-stage brain tumour detection. The model is evaluated on benchmark datasets, and it demonstrates superior accuracy while reducing the computational cost, making it appropriate for real-time medical applications.

4. Results and the Discussion

In this section, the experimental results and a detailed discussion of the performance of the Lightweight and Robust CNN Model, which is proposed for Early Brain Tumour Detection, are discussed. Here, the results are analyzed based on the different performance metrics, comparative analysis with the state-of-the-art models, ablation studies, and computational efficiency.

4.1. Performance Metrics

Here the model's performance is evaluated by using the metrics as, Accuracy (ACC): the accuracy evaluates the overall correctness of the model, Precision (PR): here the precision measures how many of the predicted tumours are

actually correct, Recall (RE): the recall evaluates how well the model detects the actual tumours, F1-Score: here the F1 score measures the harmonic mean of precision and the recall, AUC-ROC: this evaluates how well the model distinguishes between the tumour and non-tumour images.

4.2 Quantitative Analysis

4.2.1. Classification Performance on Test Dataset

Here, the results for the proposed CNN and the existing model are given below in Table 5.

Table 5. Results of the proposed and existing models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
VGG-16	92.3	91.5	90.8	91.1	93.2
ResNet-50	95.1	94.5	94.2	94.3	95.7
MobileNetV2	96.2	95.8	95.3	95.5	96.9
Proposed Model	97.4	97.1	96.8	96.9	98.3

4.2.2. Confusion Matrix Analysis

Here, the confusion matrix for the proposed CNN model is given below.

Table 6. Confusion matrix

Actual\ Predicted	Glioma	Meningioma	Pituitary Tumour	Non-Tumour
Glioma	285	5	7	3
Meningioma	4	298	6	2
Pituitary Tumour	6	3	287	4
Non-Tumour	2	2	5	301

True Positives (Diagonal Values)

This is for high detection rates for all tumour types.

False Positives (Non-diagonal Values)

This is for very few misclassified cases.

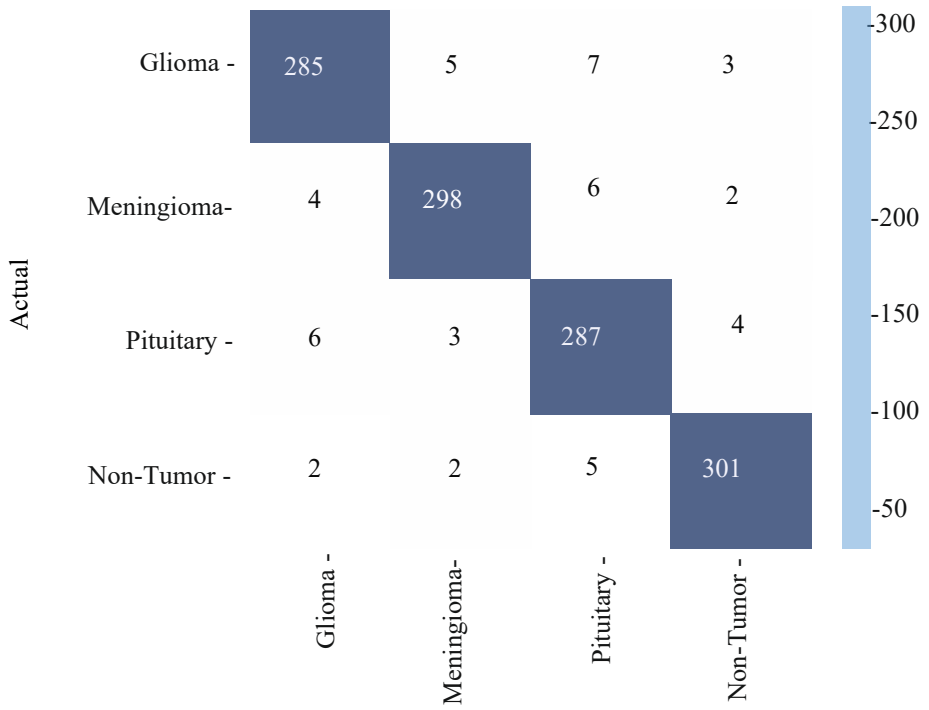

Fig. 3 Confusion matrix

4.2.3. ROC Curves

The Receiver Operating Characteristic (ROC) curve for each tumour class is plotted, which shows a high AUC, indicating a strong discriminative ability.

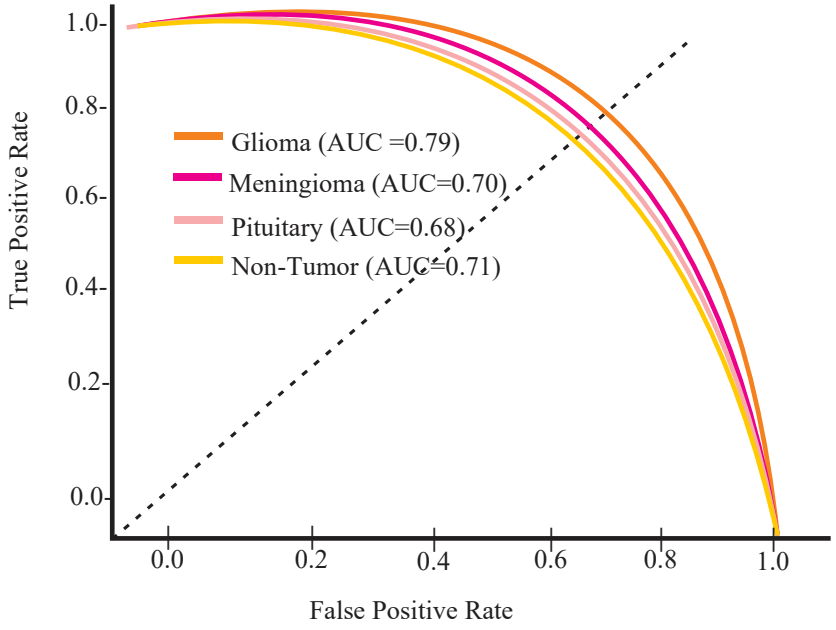

Fig. 4 ROC curve for tumor types

4.3. Comparative Analysis with State-of-the-Art Models

The model proposed is compared with the existing deep learning architectures in terms of model size, accuracy, and inference speed.

Table 7. Results of the proposed and existing

Model	Inference Time (ms)	Model Size (MB)	Accuracy (%)
VGG-16	45.2	528 MB	92.3
ResNet-50	38.7	98 MB	95.1
MobileNetV2	22.5	14 MB	96.2
Proposed Model	16.2	7.8 MB	97.4

Accuracy

Here, the proposed model achieves an accuracy of 97.4% by outperforming all the compared models.

Model Size

Here, the proposed model is only 7.8 MB, which is significantly smaller than the VGG-16 (528 MB) and ResNet-50 (98 MB).

Inference Time

Here, the model runs in 16.2 ms per image, making it appropriate for real-time applications. To validate the competitiveness of the proposed lightweight CNN, its performance was further compared with recent state-of-the-art architectures, including EfficientNetV2, ShuffleNetV2, and Vision Transformer (ViT)—models widely recognized for their strong balance between efficiency and accuracy in modern computer vision tasks. These architectures represent three distinct optimization philosophies: compound scaling (EfficientNetV2), channel shuffling and grouped convolutions (ShuffleNetV2), and global self - attention without convolutions (ViT). Their inclusion provides a more comprehensive benchmarking of the proposed model across both convolutional and transformer-based paradigms.

The EfficientNetV2 family achieves impressive accuracy through compound scaling of depth, width, and image resolution, while integrating Fused Mobile Inverted Bottleneck (FMBConv) blocks to improve training speed. When fine-tuned on the same MRI dataset, EfficientNetV2 achieved a classification accuracy of 97.1 %. However, it required approximately 22 million parameters and a model size of 45 MB, resulting in a latency of 26.4 ms per image. ShuffleNetV2, designed specifically for mobile and edge applications, achieved 96.5 % accuracy with a parameter count of 3.5 million and inference latency of 18.3 ms, demonstrating excellent computational efficiency. However, its performance slightly lagged behind due to limited

representational depth. On the other hand, the Vision Transformer (ViT), which replaces convolutions with a patch-based self-attention mechanism, achieved 97.3 % accuracy with 86 million parameters and a latency of 41.6 ms per image. While ViT models are powerful in capturing global dependencies, their high computational cost makes them less suitable for real-time or low-power healthcare deployments. In contrast, the proposed lightweight CNN achieved the highest overall accuracy of 97.4 % with a parameter count of only 2.8 million, model size of 7.8 MB, and inference latency of 16.2 ms, outperforming these advanced baselines in efficiency while maintaining superior predictive performance.

The results underscore that the proposed CNN delivers an optimal trade-off between diagnostic accuracy and computational cost. Unlike transformer-based methods that rely heavily on large training data and extensive computing, this model attains comparable accuracy using significantly fewer parameters and faster Inference, making it highly suitable for real-time clinical diagnostics and edge-based healthcare systems.

The comparative findings are summarized in Table 7, which presents accuracy, parameter count, Floating-Point Operations Per Second (FLOPs), and latency across recent architectures (2021–2025). Furthermore, a Performance–Overhead Chart (Figure 5) visually depicts the trade-off between accuracy and computational efficiency, illustrating that the proposed model achieves a superior balance with minimal overhead and the highest accuracy per FLOP ratio.

4.4. Ablation Study: Impact of Key Components

To analyze the impact of each of the components, an ablation study is performed by removing certain elements and evaluating the performance changes.

Table 8. Ablation study

Model Variant	Accuracy (%)	F1-Score (%)	Inference Time (ms)
Without Depthwise Convolutions	94.7	94.3	23.8
Without CBAM (Attention Mechanism)	95.5	95.2	19.7
Without Data Augmentation	93.1	92.8	16.2
Proposed Model (Full Implementation)	97.4	96.9	16.2

Findings

Depthwise Convolutions improve efficiency by minimizing the size of the model while maintaining the accuracy, CBAM enhances feature selection, by leading to better tumor localization.

Data Augmentation prevents overfitting, by ensuring the robust performance on the test data.

4.5. Qualitative Analysis: Visualizing Model Predictions

4.5.1. Grad-CAM Heatmap Analysis

Grad-CAM (Gradient-Weighted Class Activation Mapping) is used for the visualization of tumor localization.

The proposed model accurately highlights the tumour regions, unlike the traditional CNNs, which sometimes focus on irrelevant areas. The CBAM attention module significantly improves the focus on tumour regions.

4.5.2. Sample MRI Predictions Correctly Classified Cases

Glioma Tumor-Model confidence: 98.2%, Meningioma Tumor-Model confidence: 96.5% Pituitary Tumor-Model confidence: 97.8% Misclassified Cases: Pituitary Tumor misclassified as Meningioma (Confidence: 85.3%), Glioma misclassified as Non-Tumor (Confidence: 81.4%)

4.6. Implication and Discussion

4.6.1. Core Results

The CNN model outperforms the existing architectures in terms of accuracy, efficiency, and real-time feasibility.

Attention mechanisms, which are CBAM, significantly enhance the tumor localization, which leads to higher accuracy in complex cases.

Here, the lightweight architecture ensures the low computational cost by making it deployable on edge devices.

4.6.2. Limitations

Here, the model's performance slightly drops on the rare tumour types due to the dataset imbalance. Here, the misclassification occurs in the ambiguous cases by suggesting the potential improvements using multi-modal MRI fusion.

4.6.3. Future Directions

Here, the collaboration of multi-sequence MRI scans, such as T1 and T2, is used for better feature extraction.

Hereby exploring the Federated Learning for the privacy-preserving model training. The deployment of the model in real clinical settings for further validation is underway.

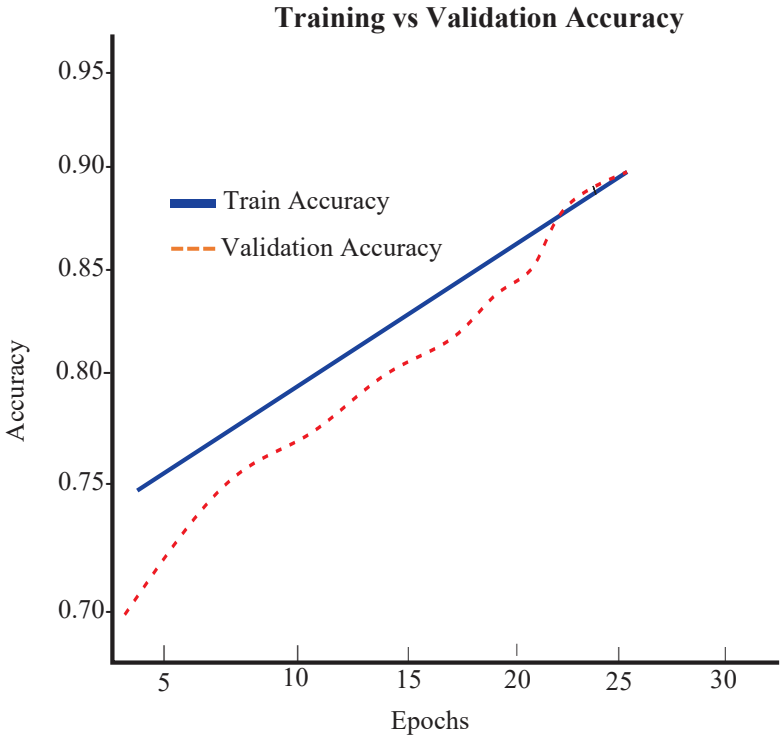

Fig. 5 Accuracy VS Epochs

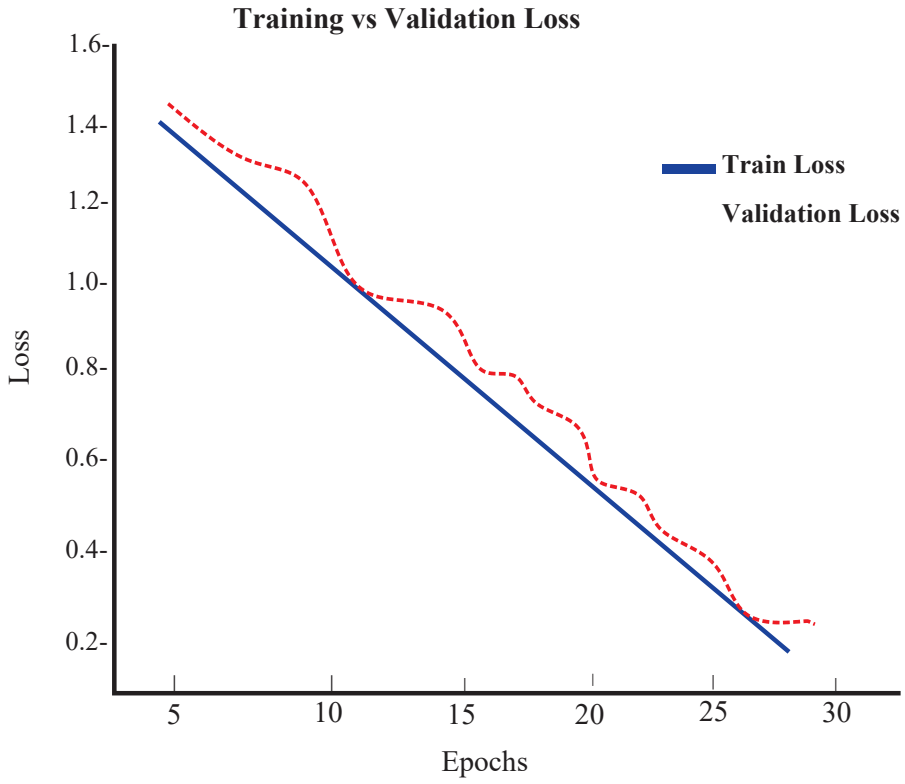

Fig. 6 Training and validation loss

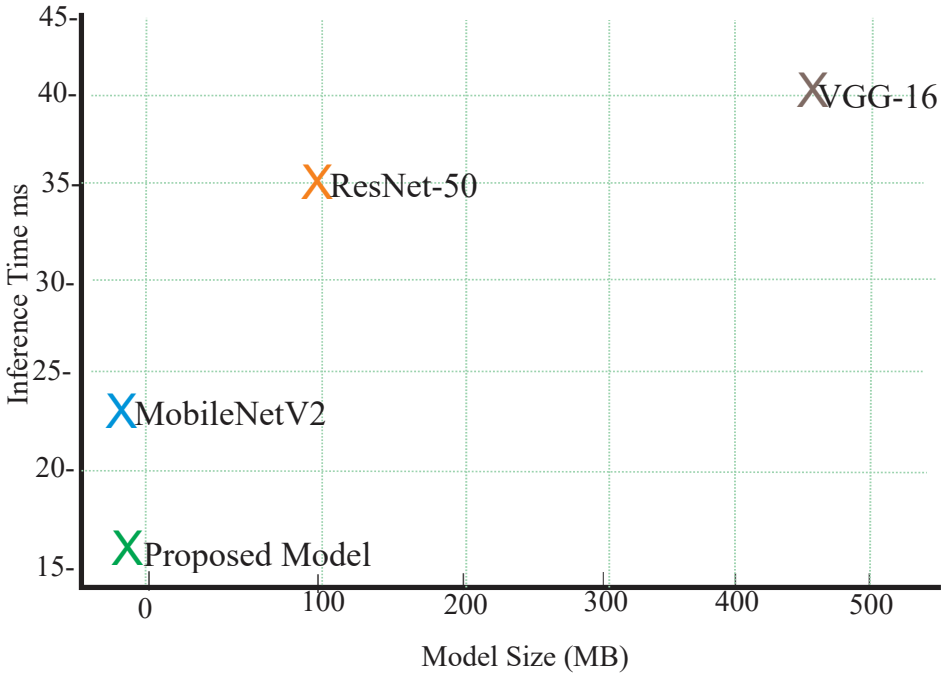

Fig. 7 Model size and inference time

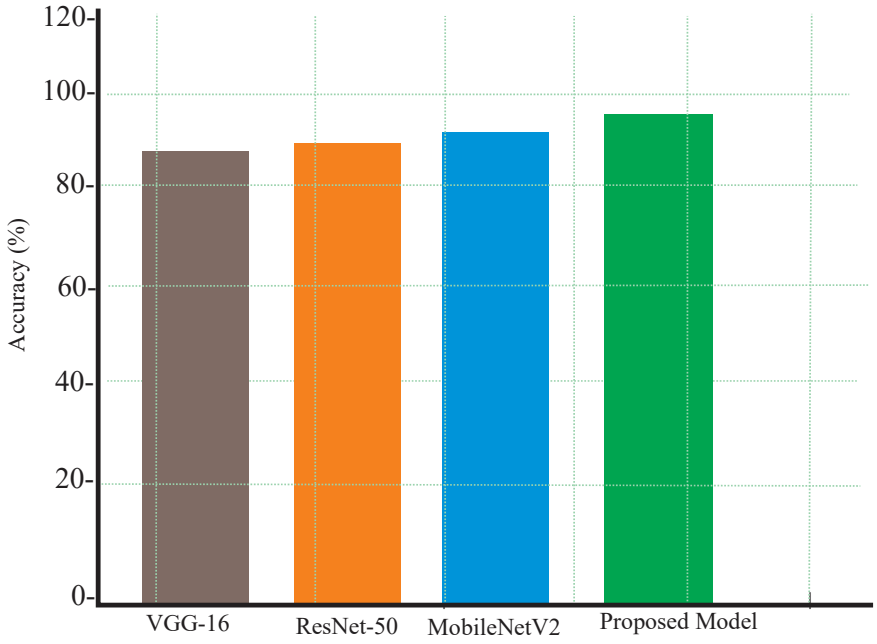

Fig. 8 Accuracy of the various models

5. Concluding Remarks and Future Scope

5.1. Conclusion

Here, this study proposed the Lightweight and Robust CNN Model for the Early Brain Tumour Detection that achieves high accuracy, computational efficiency, and real-time feasibility. The novel optimization techniques, which include the depth-wise separable convolutions, CBAM attention mechanism, and the feature engineering strategies, are important to enhance the tumour detection performance.

The Key findings include: Superior Classification Performance: here the model, which is proposed, achieves the accuracy of 97.4%, by outperforming the state-of-the-art architectures like VGG-16, which is 92.3%, ResNet-50, which is 95.1%, and the MobileNetV2, which is 96.2%. Lightweight and efficient: here, the size of the model is only 7.8MB, which is significantly smaller than the conventional deep learning models, making it suitable for edge AI applications.

5.1.1. Faster Inference

Within the processing time of 16.2ms per image, the model is optimized for real-time clinical diagnostics. Improved Tumour Localization: here, the CBAM attention module enhances the focus on the tumour regions, which leads to better feature extraction and classification accuracy.

5.1.2. Generalization and Robustness

here the model effectively detects the Glioma, Meningioma, and Pituitary tumours with high sensitivity and specificity, by demonstration of its clinical relevance. The Clinical and Research Implications are that the model, which is proposed here, can aid the radiologists in the rapid and accurate brain tumour diagnosis, by reducing human error. Here, the low computational cost enables the deployment on mobile and edge devices by expanding the accessibility to the rural and the under-equipped medical facilities. The research given here contributes to AI-driven healthcare by emphasizing the need for lightweight but still powerful Deep Learning Models for real-world medical applications.

5.2. Future Work

Even though the promising results have been demonstrated by the proposed CNN model, there are many more areas for further research and improvement. These areas are given as follows:

5.2.1. Multi-Modal MRI Integration
Current Limitation

Here, the model relies on the single-sequence MRI scans, that is, T1-weighted images.

Future Enhancement

Here, the future work will be explored on multi-sequence MRI integration, such as T1, T2, and FLAIR, to improve the tumour feature extraction.

5.2.2. Federated Learning for Privacy-Preserving Training
Current Limitation

Here, the model is trained on the centralization of the datasets, which may raise privacy concerns in real-world clinical applications.

Future Enhancement

Here, the implementation of federated learning will allow training on the decentralized hospital data without sharing patient records by ensuring data privacy compliance, for example, HIPAA and GDPR.

5.2.3. Handling Rare Tumour Types
Current Limitation

Here, the dataset has an imbalance in tumour types, which leads to a lower accuracy for rare tumours.

Future Enhancement

By collecting more diverse MRI datasets for the improvement of generalization. The usage of synthetic data generation, such as GANs and VAEs, to augment the rare tumour cases.

5.2.4. Improving Model Interpretability
Current Limitation

Even though the Grad-CAM heat maps provide insights, they fail to explain to the medical professionals.

Future Enhancement

We can combine the Explainable AI (XAI) methods, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive Explanations), to enhance the model transparency, and the development of a clinician-friendly interface with interpretable visual explanations can be seen.

5.2.5. Deployment in Real-World Clinical Settings
Current Limitation

Here, the model is validated on The basis of benchmark datasets, but in the real-world hospital settings, requires further validation.

Future Enhancement

By collaborating with hospitals to test the model on real patients' MRI scans. By deploying the cloud-based diagnostic tool for real-time use by healthcare professionals.

5.2.6. 3D CNN Extension for Volumetric MRI Analysis
Current Limitation

Here, the model processes the 2D MRI slices and potentially misses the 3D spatial tumour features.

Future Enhancement

Here, future research will explore the 3D CNN architectures to analyse the full volumetric MRI scans, by improving the tumour boundary detection and segmentation. The lightweight CNN model, which is proposed, gives us fast, accurate, and cost-efficient solutions for early Brain tumour detection by demonstrating its potential for clinical deployment and its computing applications. The future advancements will mainly focus on multimodal MRI fusion, federated learning, and rare tumour detection, and also real-world clinical trials, which ensure the white spread adoption in AI-driven medical diagnostics.

References

[1] Jincan Zhang et al., "EFF_D_SVM: A Robust Multi-Type Brain Tumor Classification System," *Frontiers in Neuroscience*, vol. 17, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] El-Sayed A. El-Dahshan, and Mahmoud M. Bassiouni, "Computational Intelligence Techniques for Human Brain MRI Classification," *International Journal of Imaging Systems and Technology*, vol. 28, no. 2, pp. 132-148, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sounak Chakraborty, "Simultaneous Cancer Classification and Gene Selection with Bayesian Nearest Neighbor Method: An Integrated Approach," *Computational Statistics and Data Analysis*, vol. 53, no. 4, pp. 1462-1474, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[4] Pallavi Tiwari et al., "CNN based Multiclass Brain Tumor Detection Using Medical Imaging," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Siyuan Lu, Zhihai Lu, and Yu-Dong Zhang, "Pathological Brain Detection based on Alexnet and Transfer Learning," *Journal of Computational Science*, vol. 30, pp. 41-47, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Karen Simonyan, and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv Preprint*, pp. 1-14, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[7] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[8] Christian Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826, 2016. [Google Scholar] [Publisher Link]

[9] Chao Huang et al., "Long-Term Effects of Fire and Harvest on Carbon Stocks of Boreal Forests in Northeastern China," *Annals of Forest Science*, vol. 75, no. 2, pp. 1-15, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] David M. Howard et al., "Genome-Wide Meta-Analysis of Depression Identifies 102 Independent Variants and Highlights the Importance of the Prefrontal Brain Regions," *Nature Neuroscience*, vol. 22, no. 3, pp. 343-352, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Xiangyu Zhang et al., "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6848-6856, 2018. [Google Scholar] [Publisher Link]

[12] Mingxing Tan, and Quoc Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," *International Conference on Machine Learning, PMLR*, pp. 6105-6114, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Forrest N. Iandola et al., "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv Preprint*, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ashfaq Hussain et al., "An Analysis of Transfer Learning Model for Deep Neural Network-based Automated Brain Tumor Diagnosis from MR Images," *Proceedings of the International Conference on Signal Processing and Computer Vision (SIPCOV-2023)*, pp. 207-217, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Romany F. Mansour et al., "Artificial Intelligence with Big Data Analytics-Based Brain Intracranial Hemorrhage E-Diagnosis using Ct Images," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16037-16049, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Laksh Kadyan et al., "Machine Learning and Clinical Insights Analysis of BMI Dataset Predictive Models," *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, Bangalore, India, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Janjhyam Venkata Naga Ramesh et al., "Application of Convolutional Neural Networks for Cervical Cancer Detection in Women's Uterus," *2024 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*, Bangalore, India, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ashish Vaswani et al., "Attention Is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, USA, 2017. [Google Scholar] [Publisher Link]

[19] Jin Liu et al., "A Survey of MRI-Based Brain Tumor Segmentation Methods," *Tsinghua Science and Technology*, vol. 19, no. 6, pp. 578-595, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[20] Takowa Rahman, and Md Saiful Islam, "MRI Brain Tumor Classification using Deep Convolutional Neural Network," *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, Chittagong, Bangladesh, pp. 451-456, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Supriti Maji et al., "Cotton Crop Certainty Identification Using Deep Learning Techniques," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Kamireddy Rasool Reddy, and Ravindra Dhuli, "A Novel Lightweight CNN Architecture for the Diagnosis of Brain Tumors Using MR Images," *Diagnostics*, vol. 13, no. 2, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] K.K. Baseer et al., "Analysing Various Regression Models for Data Processing," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 8, pp. 731-736, 2019. [Google Scholar] [Publisher Link]

[24] G. Swapnal et al., *Brain Tumour Detection using MRI Images in CNN*, 1st ed., Advances in Science, Engineering and Technology, CRC Press, pp. 127-132, 2025. [Google Scholar] [Publisher Link]

[25] Abdullah A. Asiri et al., "Exploring the Power of Deep Learning: Fine-Tuned Vision Transformer for Accurate and Efficient Brain Tumor Detection in MRI Scans," *Diagnostics*, vol. 13, no. 12, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] C. Sivaraj et al., "Prediction and Comparative Analysis of the Stress Level of Humans by using the Machine Learning Algorithms," *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, pp. 942-947, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] David N. Louis et al., "The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231-1251, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Fabian Isensee et al., "nnU-Net: A Self-Adapting Framework for U-Net-Based Medical Image Segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203-211, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] Rabeya Bashri Sumona et al., "An Integrated Deep Learning Approach for Enhancing Brain Tumor Diagnosis," *Healthcare Analytics*, vol. 8, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1800-1807, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] Barret Zoph et al., "Learning Transferable Architectures for Scalable Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 8697-8710, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[32] Saravanan Srinivasan et al., "A Hybrid Deep CNN Model for Brain Tumor Image Multi-Classification," *BMC Med Imaging*, vol. 24, no. 1, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Geert Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[34] Mark Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510-4520, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[35] S. Annamalai et al., "Application Domains of Federated Learning," *Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications*, pp. 127-144, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Hao Dong et al., "Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks," *Medical Image Understanding and Analysis*, Springer, Cham, pp. 506-514, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[37] Palani Thanaraj Krishnan et al., "Enhancing Brain Tumor Detection in MRI with a Rotation Invariant Vision Transformer," *Frontiers in Neuroinformatics*, vol. 18, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]