International Journal of Computer Theory and Engineering, Vol. 18, No. 1, 2026

Optimizing Medical MRI Brain Image Classification through Compression Analysis on Deep learning Models with Light Weight Implementation

Neena K A* and Anil Kumar M N

Department of Electronics and Communication Engineering, Federal Institute of Science and Technology,
APJ Abdul Kalam Technological University, Kerala, India
Email: neenaanzar@gmail.com (N.K.A.); mn_anilkumar@fisat.ac.in (A.K.M.N.)
*Corresponding author

Manuscript received March 17, 2025; revised May 7, 2025; accepted January 4, 2026; published March 17, 2026

***Abstract*—Medical imaging is essential for diagnosing neurological diseases. Advances in Deep Learning (DL) have greatly improved brain Magnetic Resonance Imaging (MRI) classification, enabling more accurate anomaly detection. However, the high computational and memory demands of DL models pose challenges for deployment on resource-constrained platforms such as portable medical devices and edge computing systems. This study aims to address these limitations by reducing storage and transmission demands through deep learning-based tumour classification on compressed MRI data. JPEG2000 lossless compression was applied at varying ratios to examine its effect on classification performance. The analysis focuses on understanding how different compression levels impact the accuracy and reliability of DL models. Three deep learning architectures—Convolutional Neural Network (CNN), ResNet50, and MobileNetV2—were selected to represent baseline, high-capacity, and lightweight models, respectively, and were trained on compressed MRI datasets to classify brain images into glioma, meningioma, pituitary tumor, and no tumor. Evaluation metrics included Peak Signal-to-Noise Ratio (PSNR), entropy, and Structural Similarity Index (SSIM) for image quality, and precision, recall, accuracy, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUROC) for classification performance. The experimental results indicate that CNN and ResNet50 exhibit higher classification metrics at specific compression levels. However, MobileNetV2 consistently maintains acceptable performance across all tested compression ratios. MobileNetV2 features a lightweight architecture with minimal memory requirements, making it well-suited for deployment in point-of-care devices. The model was optimized and deployed on the NVIDIA Jetson TX2 Developer Board, enabling the development of a portable, lightweight diagnostic tool for brain tumor detection.**

***Keywords*—medical imaging, MRI brain classification, deep learning optimization, model compression, edge Artificial Intelligence (AI) deployment, lightweight neural networks**

I. Introduction

Magnetic Resonance Imaging is one of the most reliable imaging methods because it provides detailed structural and functional information about the brain. However, the correct and efficient classification of Magnetic Resonance Imaging (MRI) brain images remains a challenge because of its high dimensionality. Medical image compression techniques accelerate the archiving of huge data produced by various medical imaging modalities, including X-ray imaging, Computed Tomography (CT), MRI, and Positron Emission Tomography (PET) [1]. Compressing medical image data while minimizing the loss in classification performance is crucial in the medical field. For storage in Picture Archiving and Communication System (PACS), Digital Imaging and Communications in Medicine (DICOM) use compression techniques such as JPEG, JPEG-LS, and JPEG2000 [2, 3]. Research on 2D and 3D medical images shows that compressed images can reduce bandwidth requirements by two orders of magnitude compared to uncompressed images, effectively accelerating cloud-based medical imaging services [4]. Applications in telemedicine require the transmission of images over networks. The clinical images of the patients are sent to experts for guidance from rural regions where clinical benefits may not be accessible at all times. Transmission issues emerge while sending and getting an enormous number of clinical pictures over networks, especially in regions with lower bandwidth. To tackle these issues, medical images should be compressed effectively to reduce storage costs, retrieval times, and bandwidth requirements for image transmission [5]. The practical application of the autoencoder and convolutional autoencoder models for compression is constrained due to the large number of weights they generate, which reduces the compression ratio. The performance of image compression using shallow autoencoders was unsatisfactory due to the limited fitting capability of shallow neural networks and the constraints of hardware computing power [6]. Applying medical image compression before using Convolutional Neural Networks (CNNs) is feasible and beneficial, especially for reducing the computational and storage demands associated with processing and analyzing high-resolution medical images like CT scans or MRIs. When trained on high-quality, uncompressed images, the CNN model is robust against high compression ratios [7].

One of the most popular and secure medical imaging techniques for assessing brain tissue is Magnetic Resonance Imaging (MRI). The soft tissues of the brain are more contrasted in MRI pictures than in CT ones. MRI became popular because of its high resolution and pixel density. Compressing MRI medical images using JPEG2000 lossless preserves clinically relevant information while effectively reducing the image data size. This compression technique transforms the image into frequency domain information, reducing data redundancy. Thus, it enhances training and inference efficiency, demonstrating great potential for faster and more effective processing in medical imaging applications. The DICOM standard allows clinical institutions to use image compression techniques to minimize file sizes to alleviate the issue of excessive storage burden. Radiological societies across several nations have released

DOI: 10.7763/IJCTE.2026.V18.1389

guidelines for permissible Compression Ratios (CRs) for various medical imaging modalities [8, 9]. In radiography, CT and MRI recommended CRs range from 3 to 50 [10]. Much literature discusses the effect of medical image compression on classifier accuracy. Research on mammogram image classification [11] and histopathological images [12] has shown that convolutional neural networks demonstrate robustness against compression effects. The availability of large-scale digital MRI brain datasets makes them popular for deployment in deep-learning models. Convolutional neural network-based state-of-the-art architectures are widely used for this data analysis [13–15]. A hybrid compression method for 3D medical images was proposed by Min *et al.* [16], employing segmentation, which separates the image into various sections according to its pixel density and anatomical properties. They used a Deep Neural Network (DNN) model to generate an optimal prediction for each region, achieving 38% better compression performance than JPEG2000.

Mallick *et al.* [17] proposed that the Discrete Wavelet Analysis-Deep Neural Network (DWA-DNN) classifier shows superior accuracy and performance metrics compared to traditional classifiers on brain MRI datasets. Various studies have explored the significance of medical image compression and classification within bandlimited channels. Lahiru *et al.* [18] devised a joint learning architecture that optimizes the quantization of the JPEG2000 encoder and the deep CNN-based image classifier to improve encoding and classification accuracy. Researchers have examined various compression techniques employing both frequency and spatial domain methods and integrating them into various machine learning and deep learning architectures to identify the optimal compression ratio that maximizes classifier performance. Dimililer *et al.* [19] proposed Discrete Cosine Transforms (DCT) based X-ray image compression combining machine learning frameworks. In this technique, initially, the DCT is applied to input images and employs six machine learning algorithms. An optimal compression ratio was determined to achieve high accuracy for various X-ray images. Wang *et al.* [20] proposed a Wavelet Packet Transformation (WPT)-based compression approach. They implemented WPT reconstruction using a Convolutional Autoencoder (CAE). It is further enhanced by integrating an AutoEncoder (AE) with reduced information loss. Sabbavarapu *et al.* [21] proposed a combined region growth and Otsu's thresholding technique to segment input images into Regions of Interest initially (ROI) and non-ROI by applying Discrete Wavelet Transform (DWT) and Recurrent Neural Networks (RNN). They validated the results on MRI and CT scans. Liu *et al.* [22] proposed a machine vision-oriented 3D medical image compression method using a DNN tailored for segmentation, incorporating JPEG 2000 compression with a frequency analysis module and a mapping module, thereby presenting an end-to-end compression network. Lightweight AI-driven portable equipment is essential for real-time monitoring and analysis of medical images [23, 24]. Combining lossless image compression with MobileNetV2 results in a lightweight architecture that can be effectively implemented in portable devices.

Recent studies [25–27] have explored the integration of image compression techniques, such as JPEG2000 and wavelet-based methods, to reduce data dimensionality without compromising diagnostic quality. Concurrently, lightweight CNN architectures [28, 29] and optimization frameworks for edge deployment have emerged as viable solutions to support real-time medical inference on resource-constrained devices.

This research examines the trade-off between achieving the diagnostic accuracy of cancer MRI images through deep neural network-based classification and achieving the levels of compression ratio. The findings indicate that moderate levels of compression maintain the classification accuracy of the model, but high compression ratios impair model performance. Unlike previous works that studied compression or classification separately, the novelty of this study lies in systematically analyzing the impact of JPEG2000 compression ratios on CNN classification performance, along with practical deployment validation on an edge device platform, thereby bridging the gap between compression efficiency, diagnostic fidelity, and computational feasibility.

Several decades of research have been done on brain image characterization using contemporary deep learning approaches like CNN and conventional techniques. In this work, we first compressed 2D MRI brain images at multiple encoding rates. Then, we trained three types of CNN-based model classifiers comprising CNN, ResNet50, and MobileNetV2 to classify the images into four target classes: meningioma, glioma, pituitary, and no tumour. The performance of these classifiers was evaluated using a variety of assessment measures, and their results were reported, comparing them to one another. The MobileNetV2 balances performance accuracy and architectural simplicity, making it suitable for deployment in resource-constrained environments. The significant contributions of the research paper in question have been briefly outlined below.

- The effect of different compression ratios using JPEG2000 lossless compression on medical MRI images was analysed and evaluated based on parameters such as PSNR, Compression Ratio, and Entropy.
- Investigated the effectiveness of compression ratios in automatically classifying brain MR images into glioma, meningioma, pituitary, and no tumour using three well-known, distinct pre-trained CNN models: CNN, ResNet50, and MobileNetV2 with transfer learning.
- The lightweight MobileNetV2 model is trained and deployed on the NVIDIA JETSON TX2.

This proposed work demonstrated the potential of compression on deep convolutional network architectures pretrained with transfer learning combined and image pre-processing techniques. This work could be a basis for formulating automatic tumour classification in bandwidth-constrained channels and for real-time medical imaging applications. The manuscript was divided into five sections: Section II elaborates on the materials and methodology. Section III discusses experimental evaluation metrics. Section IV presents the experimental results and discussion. Finally, the manuscript concludes with a summary and scope for future work.

II. Materials and Methods

This section describes the step-by-step approach followed for brain MRI image classification using deep learning models. It provides information on data acquisition, preprocessing, compression methods, model selection, and deployment on the NVIDIA Jetson TX2 platform. The whole workflow is designed to be efficient in terms of both model performance and hardware-level deployment, making it appropriate for real-time diagnosis in edge environments. Fig. 1 shows the step-by-step process of classifying MRI brain images with a deep learning approach using an NVIDIA Jetson TX2 development board for edge deployment. It begins with MRI brain image acquisition and proceeds toward pre-process operations such as resizing, noise removal, and normalization for uniformity. The image augmentation techniques such as rotation, flipping, and scaling enrich the dataset and help the deep learning models generalize. We investigated the impact of image compression on an MRI brain image data set. We used an MRI dataset from Kaggle [30] containing human brain images. This data set is a combination of Figshare, Br35H and SARTAJ dataset.

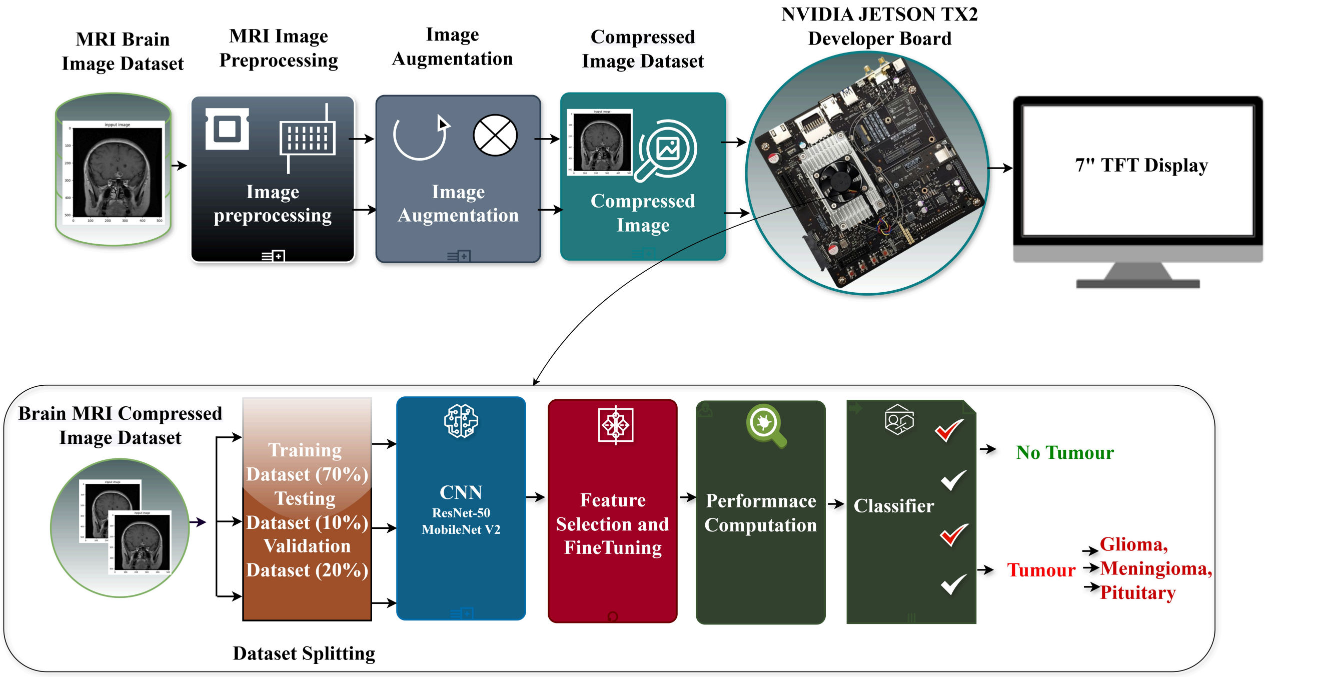

Fig. 1. The architecture of compression analysis on deep learning models and lightweight implementation deployed on an NVIDIA JETSON TX2 Developer board.

The dataset of MRI images is categorized by type of tumor as glioma, meningioma, pituitary, and no tumor which is detailed in Table 1.

Table 1. Distribution of MRI image dataset across training, validation, and testing sets

Label	Train	Validation	Test Data Set	Total
Glioma	931	400	300	1631
No Tumor	1123	481	405	2009
Meningioma	951	408	306	1665
Pituitary	1020	437	300	1757
Total	**4025**	**1726**	**1311**	**7062**

Then, the augmented images undergo compression, creating a lightweight dataset that can be deployed on the Jetson TX2 for inference and displaying real-time results on a 7-inch TFT display. The bottom part of the figure describes the deep learning process. The compressed dataset has been divided into training (70%), testing (10%), and validation (20%) sets so that the model can be trained and tested. ResNet-50, CNN and MobileNet V2 are the convolutional neural networks employed for feature extraction and classification. The extracted features were further fine-tuned to increase the classification accuracy. Various metrics have been applied to evaluate the models, especially accuracy, precision, recall, and inference speed, which are specific to the Jetson TX2 hardware. The classifier decides whether the MRI scan is tumour-free or belongs to one of the separate tumour classifications, such as glioma, meningioma, or pituitary tumour. This workflow emphasizes the deployment of real-time edge AI, lightweight model implementation, and accurate classification, making it a practical solution for medical diagnostics in resource-limited environments.

Among these three architectures, the MobileNetV2 framework was further optimized and deployed on the NVIDIA JETSON TX2 Developer board using the TensorFlow framework. It is a powerful platform for edge AI applications. It is capable of real-time medical image classification and other computationally complex tasks. It features a quad-core ARM Cortex-A57 CPU with 2 MB L2 cache, an NVIDIA Pascal GPU with 256 CUDA cores, and 8 GB of 128-bit LPDDR4 RAM, providing high-performance computing designed for embedded AI tasks. TensorFlow is an open-source light Machine Learning (ML) framework for deploying deep learning models designed for mobile and embedded devices. TensorFlow provides a library of pre-trained models and tools for model pruning, quantization, and compression, which is critically important for deploying deep learning models on edge devices. The Jetson GPU enables accelerated computations, enhancing the performance of lightweight architectures like MobileNetV2. The following sections elaborate on the image preprocessing techniques, the application of JPEG2000 image compression and the implementation details of the deep learning models used in this research.

A. Image Preprocessing

The input requirements of the deep learning models were achieved by applying pre-processing steps to the MRI brain images. These steps include random noise removal, resizing of the images, and normalization of the pixel values. As mentioned in Ref. [31], data augmentation methods such as translations, flipping, and rotations are done to artificially expand the dataset as well as mitigate the risk of overfitting during training. These methods guarantee that the images maintain high standards of quality and consistency. After pre-processing, the images are enhanced by applying JPEG2000 compression which reduces the image size while maintaining the core diagnostic features for classification.

The proposed method adopts a systematic enhancement of the image quality and the deep learning model performance. To begin with, every MRI image is resized to 224×224 pixels for compatibility with popular frameworks CNN, ResNet50, and MobileNetV2. Such MRI images can now be processed as a single batch during training without losing important features that need to be classified precisely. This standardization ensures consistent batch processing during training while preserving critical visual features needed for accurate classification. It includes random noise removal to significantly enhance SNR for better feature extraction while normalizing pixel values to a range between 0 and 1. Pixel value normalization not only increases training stability, but enhances performance and expedites convergence during model training. The combination of these preprocessing techniques along with augmentation increases the model's resilience to varied inputs for reliable classification.

A second important aspect of this methodology is JPEG2000 compression on images that have already been pre-processed. JPEG2000 compression diminishes the quantity of data, but maintains some of the primary medical qualities deemed necessary for classifying instances accurately [32, 33]. In particular, compression is beneficial for modern resource-limited scenarios, including mobile health care systems and telemedicine platforms, where every stored value and computation has to be optimally efficient. In the case of tumor boundaries and other diagnostic features, JPEG2000 maintains critical medical information while removing redundant details during pre-processing. Operations such as resizing, pre-processing, data augmentation, and compression ensure that deep learning algorithms receive homogeneous, optimal quality inputs. This technique improves the reliability and precision of brain MRI scans, reducing computational load.

B. JPEG2000

With the help of the Glymur module, version 0.13.6, in this study, JPEG-2000 compression has been done. Glymur is a Python interface to the OpenJPEG library [34]. As for JPEG, wavelet-based compression is used, as opposed to the block-based DCT method used in traditional JPEG. This results in a representation of the image, with fewer blocking artefacts and a smoother structure. The 5/3 reversible wavelet transform is a method of lossless compression in which some parts of the pixel data are retained while others are discarded.

An image is decomposed by wavelets into a set of its lower-resolution (coarse) and a number of higher-resolution (fine) representations (sub-images) in such a way that the waveforms of each of the sub-images are arranged in a different frequency range, and the sub-images are called wavelet coefficients. This works in favour of medical images, as more detail is usually required, for instance, in the texture of the tumour and the edges. The compression levels analysed in this study range from clinically acceptable ratios, below 50:1, to significantly higher ratios. The aim here is to analyse the point at which important detail and features will no longer be preserved.

Investigating how compression impacts deep learning classifications and finding tipping points that still retain critical information prompted this test. Additionally, this aligns with practical concerns in storage and transmission, where the minimization of image data is of utmost importance. Overall, JPEG2000 compression provides balanced optimization with sufficient quality and the necessary attributes for dependable AI-based cancer detection in images.

Fig. 2 shows the compressed image at various Compression Ratios (CRs) using a patch extracted from a sample MRI brain image. The figure includes 13 different compression levels, each represented by a subfigure. Each subfigure displays the selected patch alongside its respective compression ratio. An increase in CR leads to a noticeable degradation in image quality. This demonstrates the direct correlation between compression ratios and image clarity.

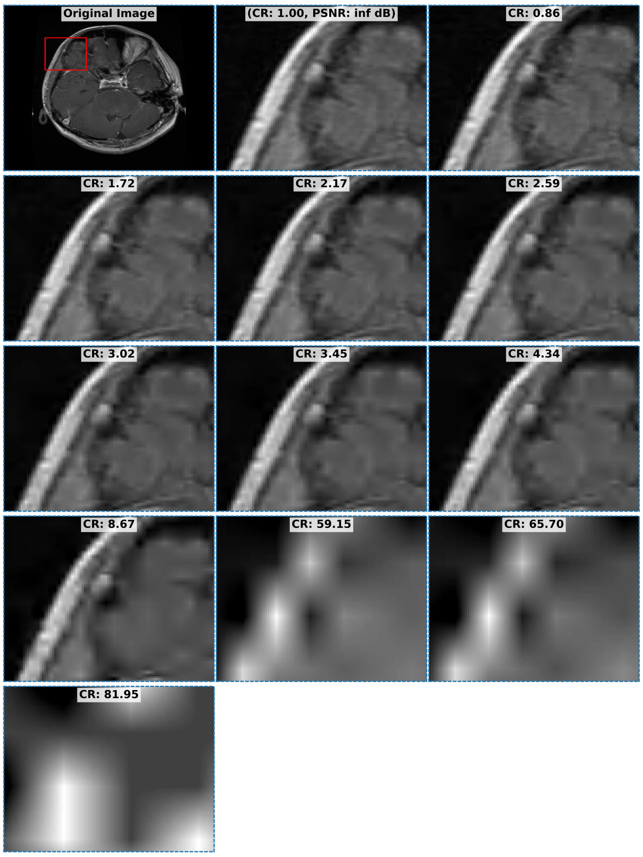

Fig. 2. Compressed images at various compression ratios.

C. Deep Learning Models

In this research, we trained three types of deep learning architectures, CNN, ResNet50, and MobileNetV2, using pre-processed and compressed data. These architectures were evaluated with compressed images at varying levels of compression. For each model, training was conducted across thirteen different compression ratios to thoroughly analyze their performance and adaptability to compressed inputs. The following subsections provide a detailed description of the architectures.

1) Convolutional Neural Network (CNN)

A CNN architecture extracts hierarchical features from input image data using multi-layer neural networks. CNN is a widely used and popular method for medical image classification [35]. In this research, CNN is designed to classify brain image data set into four classes using the TensorFlow framework. Fig. 3 depicts the architecture of the Convolutional Neural Network (CNN) for brain tumor classification using RGB-converted MRI scans. This architecture consists of multiple convolutional, pooling, dropout, and fully connected layers. Input images of dimensions 224×224×3 are processed through four convolutional layers with 3×3 kernels and ReLU activation functions. Pooling layers were added to condense the information from the convolution layer. A Max pooling layer is added to reduce spatial dimensions while keeping significant features. The feature depth increases progressively from 32 to 256 channels, with ReLU activation and kernel regularization (L2 regularization, $\lambda = 0.001$), enabling multi-scale feature representation. Overfitting is reduced by adding dropout layers of rates 0.2 and 0.5. Flattening layers are added further. The model is compiled with the Adam optimizer set to a learning rate of 0.0001. The training was conducted over 50 epochs.

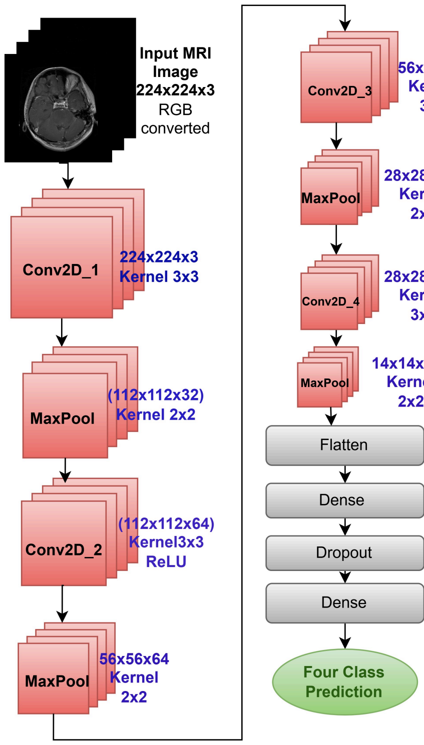

Fig. 3. Convolutional Neural Network (CNN) architecture for brain tumor classification from MRI images.

2) ResNet50

ResNet-50 is a Convolutional Neural Network (CNN) for efficient image classification tasks. It is a medium-sized model in the Residual Networks family [36]. It is designed to overcome the difficulties and complexities associated with training in deep neural networks. ResNet learns residual functions in each input layer. This work develops a deep learning framework by leveraging the benefits of transfer learning and adding more neural network layers for multi-class image classification. Fig. 4 illustrates the architecture of ResNet50 employed for classifying MRI brain images. The model is initialized with pre-trained ImageNet weights to enhance feature extraction capabilities and is modified by freezing the original fully connected top layers, followed by the addition of task-specific classification layers. In ResNet50 architecture, the added custom layers are flattened, dense layers of 512 and 32 neurons with ReLU activation function and dropout layers of rate 0.2. Training is performed using the Adam optimizer and the categorical cross-entropy loss. ReduceLROnPlateau has been added to adapt the learning rate according to the accuracy of the validation. A batch size of 64 and 50 epochs are used for model training. This approach combines the robustness of pre-trained models with the flexibility of custom layers and achieves efficient feature learning.

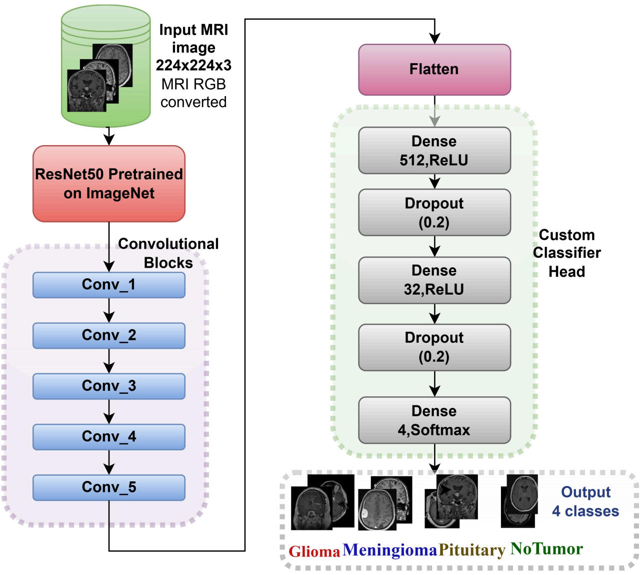

Fig. 4. ResNet50 architecture for MRI brain image classification.

3) MobileNetV2

MobileNetV2 [37] is a convolutional neural network framework that enhances performance on mobile and limited-resource devices. It features an inverted residual design with linear bottlenecks, where the input and output of the residual block comprise narrow bottleneck layers. The intermediate expansion layer employs lightweight depth-wise convolutions to extract and process features from the input image. Across various applications, including object detection, semantic segmentation, and image classification, this architecture maintains good accuracy while drastically reducing computational complexity [38–40]. The architecture is popularly deployed in deep learning models on mobile platforms. Fig. 5 presents a modified MobileNetV2 architecture designed for classifying brain MRI images into four categories: Glioma, Meningioma, Pituitary, and No Tumor. The input images are resized to

224×224×3 and passed through an initial convolutional layer, followed by a series of depthwise separable convolutional bottleneck blocks with ReLU6 activation. These lightweight bottleneck layers progressively extract rich features while maintaining computational efficiency. The architecture incorporates global average pooling to reduce spatial dimensions before passing through dense layers, including dropout regularization. A final softmax output layer predicts class probabilities for the four tumor types, making the model suitable for deployment on resource-constrained medical devices.

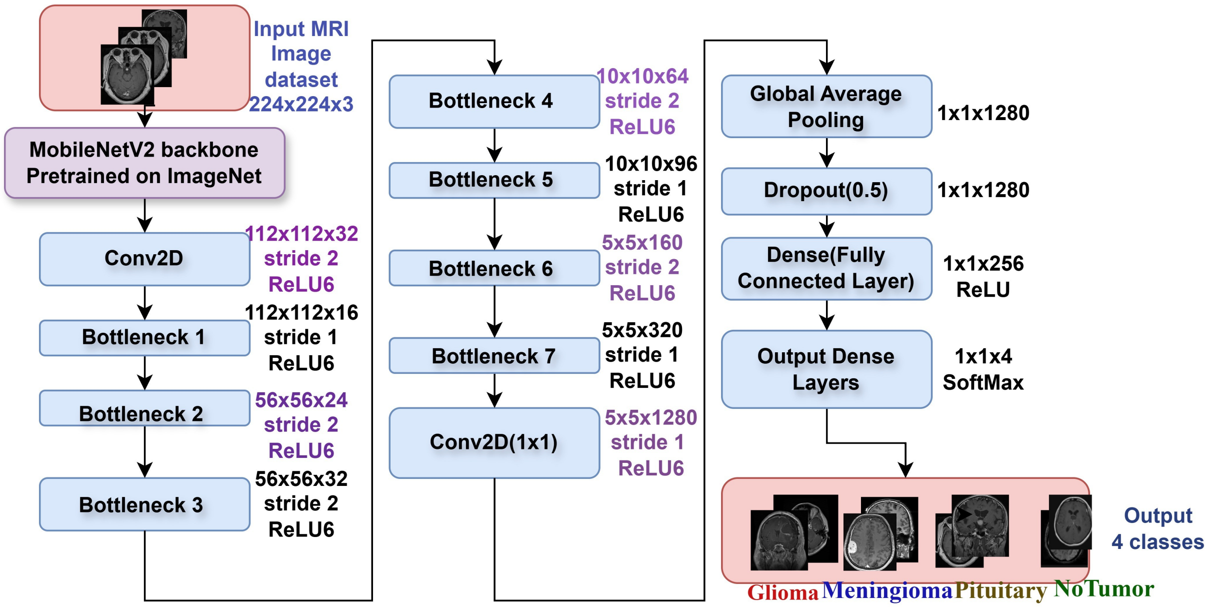

Fig. 5. MobileNetV2 architecture for brain tumor classification from MRI images.

III. Evaluation Metrics

The experiment was carried out to evaluate various degrees of Compression Ratios (CRs) to measure their impact on image quality using metrics such as the Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and entropy. The different compression ratios employed in this analysis proceed from the minimum to the maximum, illustrating that the increased compression affects the image fidelity to some degree. The performance metrics are calculated using the equations described below.

1. SSIM: SSIM is a commonly used metric to evaluate the perceptual similarity between images, specifically the original x and the compressed image y. The formula for SSIM is given by Eq. (1).

$$\text{SSIM}\ (x, y\ =\ [l\ (x, y)]^{\alpha}\ \cdot\ [c\ (x, y)]^{\beta}\ \cdot\ [s\ (x, y)]^{\gamma} \quad (1)$$

where x and y are the two images being compared. The brightness, contrast, and structural components are represented by the variables l, c and s, respectively. The exponents α, β, and γ modify the relative significance of various elements.

2. CR: The compression ratio quantifies the size reduction achieved during compression. It is calculated using Eq. (2).

$$CR\ =\ \frac{Original\ File\ Size}{Compressed\ File\ Size} \quad (2)$$

3. PSNR: PSNR measures the quality of the compressed image for the original image. It is expressed in Eq. (3).

$$PSNR\ =\ 10\ log10\,\frac{IMax^2}{MSE} \quad (3)$$

where $IMax$ is the maximum pixel value of the image (e.g. $Imax$ = 255 for an 8-bit image), and the Mean Squared Error (MSE) is given by Eq. (4):

$$MSE\ =\ \frac{1}{MN}\sum_{i=0}^{M}\ \sum_{j=0}^{N}\ |I(i,j)\ -\ I\_com(i,j)| \quad (4)$$

where, M and N are the dimensions of the images in height and width, respectively.

4. Entropy (H): Entropy measures the information content in the compressed image. It is computed using the Eq. (5).

$$H(X) = -\sum_{i=0}^{n} P(Xi)\cdot log_2(P(Xi)) \quad (5)$$

where, $P(Xi)$ is the probability of each intensity value occurring in the image histogram, and n is the total number of unique intensity levels.

Models developed for classification tasks are assessed using performance evaluation metrics like accuracy, precision, recall, and F1 score. These methods are obtained from the confusion matrix.

1. Accuracy: The classifier's accuracy can be determined by calculating the ratio of correct predictions to the total number of predictions, as shown in Eq. (6).

$$\text{Accuracy}\ =\ \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (6)$$

In the above equation, True Positives (TP) and True Negatives (TN) represent correctly classified positive and negative samples respectively. False Positives (FP) and False Negatives (FN) represent incorrectly classified samples.

2. Precision: It refers to the number of correctly predicted positive classes by the model out of the total number of true

positive classes. The precision is given in Eq. (7) below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+ FP}} \tag{7}$$

3. Recall: Recall measures how accurately our model predicted the positive classes from the total number of positive classes and is given by Eq. (8).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

4. F1 score: The F1 score is a performance parameter used to measure predictive performance. The F1 score can be determined using the formula in Eq. (9).

$$\text{F1 score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{9}$$

5. The Area Under the Curve (AUROC): It is a metric used to represent model performance in a single metric [41]. The Area Under the Curve (AUC) values lie in the range 0 to 1. The value of AUROC of a good classifier is close to one. A higher AUC value frequently indicates a better model performance.

IV. Result and Discussion

The experimental results are demonstrated to evaluate the impact of compression on medical MRI images and the performance of deep learning architectures. The following sections provide a detailed analysis, starting with the effect of image compression and ending with an evaluation of deep learning models.

A. Image Compression

Table 2 presents a comprehensive evaluation of JPEG2000-compressed brain MRI images across a wide range of compression ratios, using four key metrics: SSIM, PSNR, Entropy, and Compression Time. These indicators collectively help in understanding how compression affects both image quality and processing efficiency, which is especially crucial in medical contexts.

SSIM reflects how closely the compressed image retains structural features of the original, including textures, edges, and fine details. As shown in the table, SSIM values gradually decrease as compression becomes more aggressive. For compression ratios below 6:1, SSIM remains consistently high (above 0.9), indicating near-lossless quality suitable for clinical diagnosis. However, once the compression ratio exceeds 25:1, SSIM drops significantly, falling to as low as 0.515 at 81.94:1, implying noticeable degradation in image structure and diagnostic quality.

PSNR is another widely accepted measure of image quality, indicating how much noise is introduced during compression. The PSNR values show a similar downward trend with increasing compression, starting from 43.5 dB at lower ratios and dropping below 30 dB at 81.94:1. Typically, PSNR values above 30 dB are considered acceptable in medical imaging. Ratios above 25:1 often fall below this threshold, suggesting a compromise in visual fidelity that may not be suitable for high-precision tasks.

Entropy measures the amount of information or complexity in an image, shows a relatively stable trend across different compression levels. The values increase slightly and peak around 41.88:1 (5.155), suggesting that some structural complexity is preserved even at higher compression levels. This indicates that although compression simplifies data, the essential informational content isn't entirely lost, which is promising for downstream analysis or AI-based processing.

Table 2. Performance metrics evaluated for different compression ratios

Compression Ratio	SSIM	PSNR	Entropy
1:1	1	inf	4.919
1.72:1	0.982	43.496	4.974
2.58:1	0.971	40.942	4.995
3.45:1	0.961	39.276	5.012
4.33:1	0.946	38.241	5.051
6.42:1	0.925	36.543	5.038
8.67:1	0.893	35.637	4.921
12.78:1	0.868	34.851	4.932
16.99:1	0.843	34.295	4.933
25.35:1	0.788	33.531	4.993
33.99:1	0.757	33.216	4.969
41.88:1	0.704	32.455	5.155
59.14:1	0.656	32.053	5.149
81.94:1	0.515	29.828	5.07

Overall, the findings suggest that compression ratios up to 6:1 offer the best balance between preserving critical diagnostic information (as indicated by high SSIM and PSNR) and achieving meaningful reductions in file size and transmission time. Ratios higher than 25:1 introduce substantial degradation in image quality and should be used judiciously, particularly when diagnostic integrity is paramount. However, for non-diagnostic tasks such as archival or telemedicine in bandwidth-limited settings, higher compression may still be acceptable.

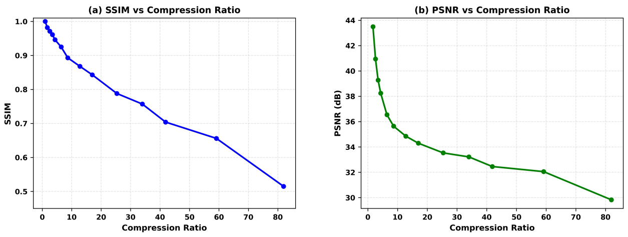

Fig. 6. Visual analysis of the impact of JPEG2000 compression on MRI image quality and performance: (a) SSIM vs. compression ratio; (b) PSNR vs. compression ratio.

Fig. 6 illustrates the relationship graph between compression ratios and three evaluation metrics: SSIM and PSNR. The *x*-axis represents the compression ratio, and the *y*-axis shows the corresponding metric values. Overall, the graph highlights how compression affects image quality differently across metrics, with PSNR being the most sensitive to increasing compression ratios.

In Fig. 6, data size reduction is improved, but at the cost of image quality by increasing compression ratios. This reduction in image quality occurs due to the information loss during the compression stage. Compression using JPEG2000 leads to the degrading of high-frequency information and texture content in the image. Considering this factor, it is essential to maintain a trade-off between medical image quality and compression. A compression ratio limited to ≤6.42:1 maintains good diagnostic quality; therefore, it is suitable for high-accuracy tasks. Moderate compression ratios (8.67 to 16.99) can be used for general-purpose or storage purposes. Higher compression ratios (≥25.35) may be acceptable for archival and secondary analysis, considering the quality and file size trade-offs. The compression ratio below 6:1 results in minimal loss of quality and is clinically acceptable. The compression ratios between 6:1 and 25:1 balance file size reduction with diagnostic integrity due to SSIM values higher than 0.8 and PSNR values at 35 to 40 dB. Compression ratios greater than 25:1 suffer from a high loss in quality and are not ideal for diagnosis.

B. Deep Learning Models

CNN, ResNet50, and MobileNetV2 models are designed to extract meaningful features of images, even when compressed. Since these models are pretrained, the feature extraction capabilities are well generalized, effectively compensating for the loss of fine-grained details caused by compression. Architectures trained with compressed images will likely adapt to the artefacts introduced by similar compression ratios, thereby maintaining performance. Clinically relevant features, such as tumour boundaries and textures, must be retained for image classification even at high compression ratios. MobileNetV2 uses depth-wise separable convolutions, which are computationally efficient; however, there is a slight decrease in accuracy at higher compression ratios compared with CNN and ResNet50, but it remains robust across compression ratios.

Table 3 provides an approximate memory usage of the classifiers. The first column indicates the approximate memory required to store the model's trainable weights. The second column shows the estimated memory required to store the model's intermediate feature maps generated during the forward pass of the model. The third column provides the overall memory required. Memory usage was calculated by estimating the memory required to store model parameters and intermediate activations during a forward pass. The table highlights the computational efficiency of MobileNetV2 (~24 MB). CNN architecture has the highest memory usage. CNN (~380) and ResNet50 (~115) consume more memory than MobileNetV2.

Table 3. Estimation of memory usage for the architectures CNN, ResNet50, and MobileNetV2

Model	Parameters (MB)	Total Memory (MB)
CNN	~360	~380
ResNet50	~100	~115
MobileNetV2	~14	~24

Table 4 presents the classification accuracy of deep learning models CNN, ResNet50 and MobileNetV2 evaluated over various compression ratios. CNN shows a slightly higher accuracy than MobileNetV2, but lower than ResNet50. ResNet50 shows high accuracy at lower compression ratios. MobileNetV2 exhibits robust accuracy across all compression ratios, indicating a balance between performance and computational efficiency. Even though accuracy is slightly lower than ResNet50, it maintains a competitive despite at moderate compression ratios. At lower compression ratios, like 1.72:1, 3.45:1, and 4.33:1, MobileNetV2 achieves accuracy close to ResNet50 and CNN. At higher compression ratios, 59.146:1, 65.69:1, 81.94:1, MobileNetV2 shows slightly less accuracy but remains within the reasonable range given the complexity of the task and high compression.

Table 4. CNN, ResNet50, and MobileNetV2 accuracy over different compression ratios

Compression	Accuracy		
Ratio	CNN	ResNet50	MobileNetV2
1:1	0.967	0.968	0.937
1.72:1	0.9645	0.965	0.95
3.45:1	0.9675	0.968	0.9445
4.33:1	0.9605	0.9865	0.949
6.42:1	0.959	0.978	0.9375
8.67:1	0.9555	0.9755	0.945
12.78:1	0.964	0.9675	0.938
16.99:1	0.941	0.955	0.9255
24.34:1	0.956	0.967	0.919
33.98:1	0.945	0.9575	0.9205
41.88:1	0.94	0.9665	0.9125
59.146:1	0.9335	0.943	0.9035
65.69:1	0.929	0.929	0.892
81.94:1	0.927	0.9175	0.898

Fig. 7 illustrates the accuracy variations of CNN, ResNet50, and MobileNetV2 across different compression ratios using a heatmap representation. Each compression ratio is represented by a group of three bars corresponding to the accuracy of three models. From the accuracy matric report, MobileNetV2 is recommended for lightweight, real-time portable devices with a compression ratio ≤12.78:1 (Accuracy 0.94). CNN is effective at compression ratios up to ≤16.99:1. ResNet50 is preferred for high-accuracy tasks, maintaining robust performance at compression ratios up to ≤41.88:1. The lightweight nature and efficiency of MobileNetV2 make it particularly well suited for real-time medical applications.

Fig. 8 illustrates the performance metrics—precision, recall, and F1 score—for CNN, ResNet50, and MobileNetV2 models across varying compression ratios, visualized as a heatmap to highlight variations in classification effectiveness under different compression levels.

From Fig. 8, it can be observed that ResNet50 demonstrates slightly superior performance in terms of metrics such as precision, recall, and F1 score of certain compression ratios. Although MobileNetV2 shows slightly lower performance metrics compared to ResNet50 and CNN, it maintains a consistent and balanced performance across varying compression ratios. Also, it shows a good classification performance even at high compression ratio values—a precision value of 0.903 at a CR of 59.146 and 0.89 at a CR of 81.94. The experimental results and heatmap show that MobileNetV2 provides a strong balance between classifier performance metrics and efficiency over different image compression ratios. While ResNet50 and CNN give the highest matric values, their computational complexity and memory requirements are high, making them less feasible for portable and embedded applications.

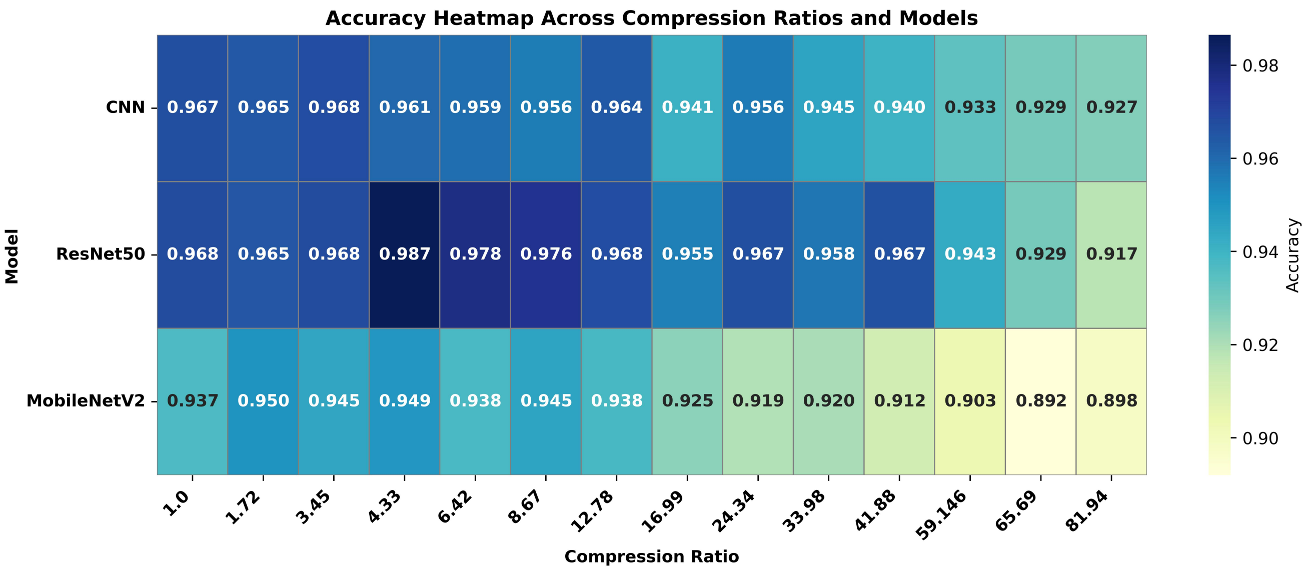

Fig. 7. Heatmap showing classification accuracy for CNN, ResNet50, and MobileNetV2 architectures across various JPEG2000 compression ratios. Darker shades indicate higher accuracy, providing a clear visual comparison of model robustness against compression effects.

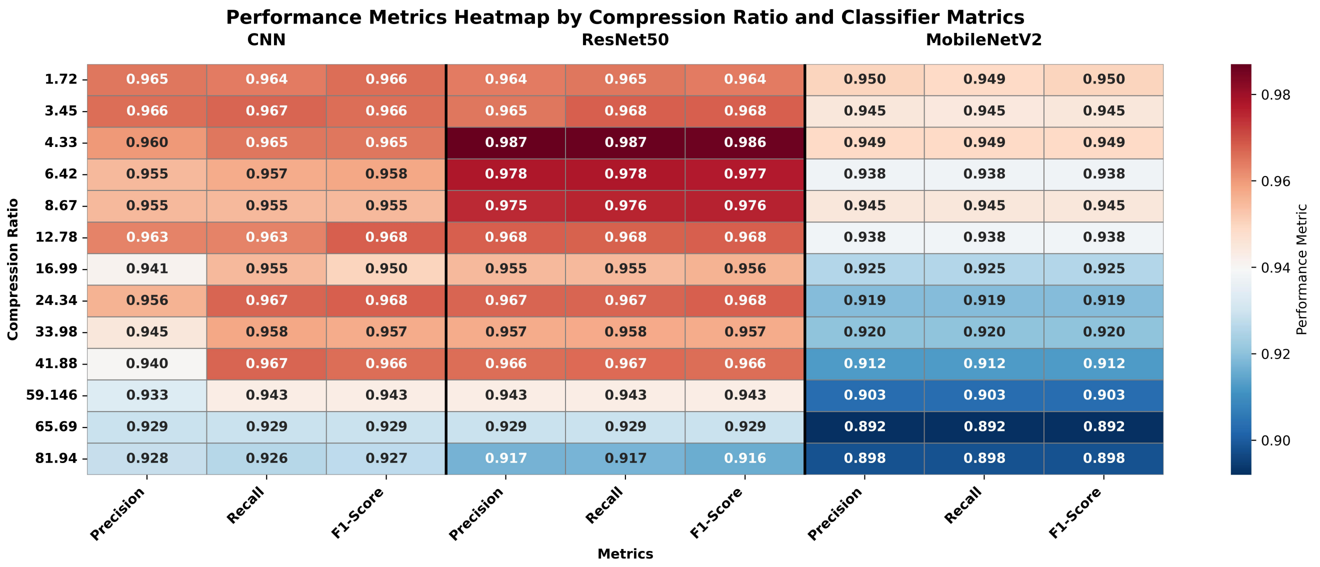

Fig. 8. The heat map performance matrices precision, recall, and F1 score for the classifiers CNN, ResNet50, and MobileNetV2 over various compression ratios.

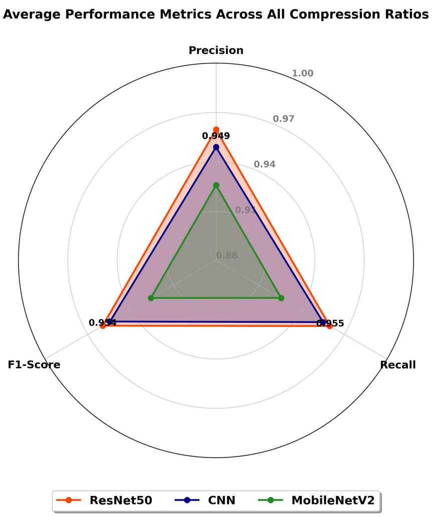

Fig. 9. Comparison of precision, recall, and F1 score averages of CNN, ResNet50, and MobileNetV2 for all compression levels presented in a radar chart. The radial axis is limited to the range [0.88, 1.00] in order to better visualize minor deviations in performance in the high-accuracy range.

To create a visual display for classification behavior for each compression ratio, a radar chart was generated, of which the mean for each Precision, Recall, and F1 score for every model is presented in Fig. 9.

Fig. 10 illustrate the average ROC AUC scores of the CNN, ResNet50, and MobileNetV2 models across varying compression ratios for the classification of glioma, meningioma, pituitary tumor, and no tumor. This figure highlights the robustness of each model in retaining classification performance under increasing compression.

All three models demonstrate high discriminative capability with ROC AUC values consistently above 0.96. ResNet50 achieves the highest ROC AUC. Meanwhile, MobileNetV2 achieves slightly lower ROC AUC values than ResNet50 and CNN. However, MobileNetV2 maintains consistent performance across all classes and compression ratios. Thus, MobileNetV2 perfectly balances classification performance and resource efficiency, making it a suitable choice for real-time medical image classification. MobileNetV2 was the chosen lightweight model for this study due to its stable performance versus compression ratios of 59.146:1 with an AUROC of 0.98 likely due to the efficiency of the model architecture. In order to assess the practicality of real-world edge of the model, we deployed it in the NVIDIA Jetson TX2 Developer Board. This model is known for its balanced power efficiency and computational capacity, making it apt for Point-of-Care diagnostic applications. For this practical implementation, we transformed the model in TensorRT and used FP16 precision to achieve lower inference latency and memory usage while maintaining the same level of accuracy the model predicted before the optimizations.

Across all models tested, the TX2 was able to infer compressed MRI images with an end-to-end latency between 12–35 ms per frame, depending on the compression level and resolution. The edge device was also able to run multiple iterations of the tests without thermal throttling, showing that the compressed-model pipeline was able to function dependably even when under the constraints of edge computing. These findings suggest that the Jetson TX2 is able to execute compressed deep models successfully in scenarios with extremely limited resources and mobility, as is the case in many facets of medical imaging, and particularly when models like MobileNetV2 and compression-methods based on wavelets are used.

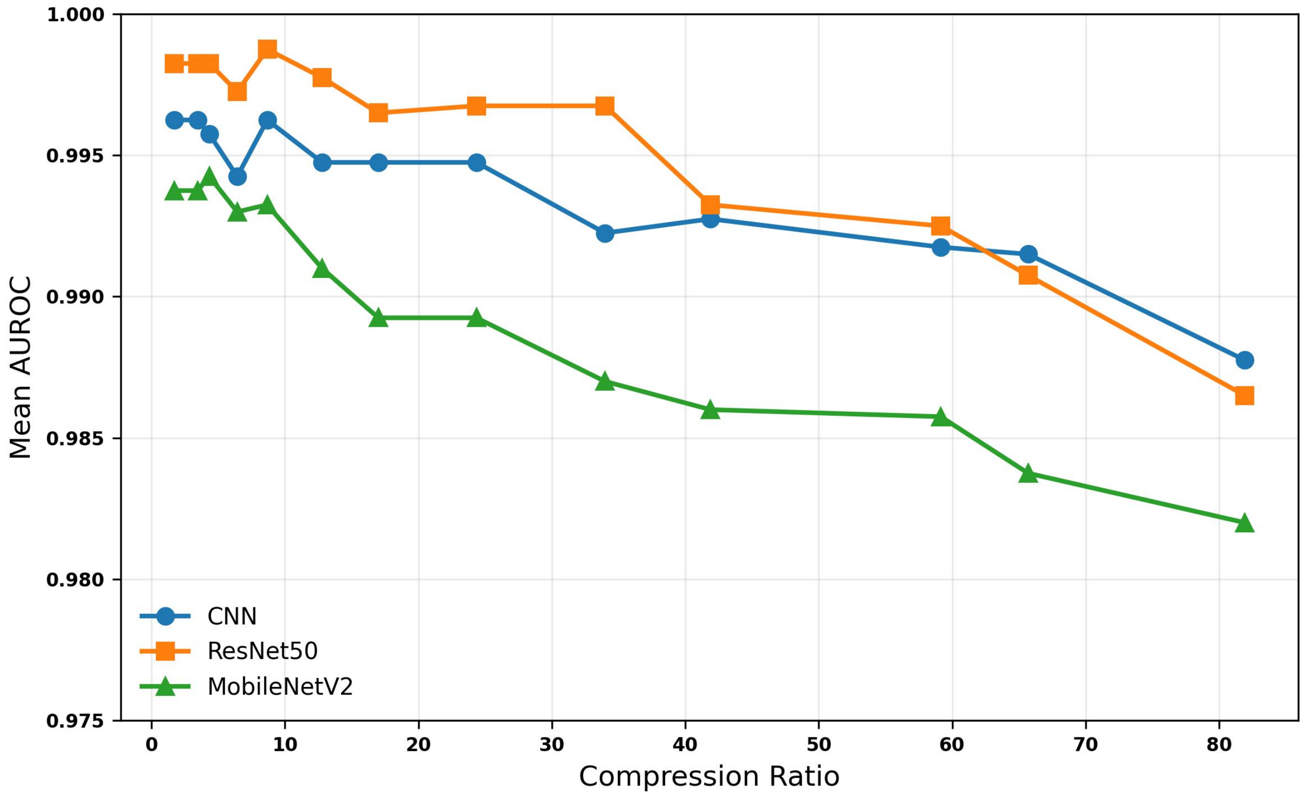

Fig. 10. Mean ROC-AUC values for the three deep learning models across tumor classes and varying compression ratios.

V. Conclusion

This study was aimed at analyzing the effect of JPEG2000 compression on brain MRI classification done with deep Convolutional Neural Networks (CNN). With respect to diagnostic accuracy, our results indicate that MobileNetV2, a lightweight model, can sustain up to 12.78:1 compression ratio with no significant loss in accuracy, thus making it viable for real-time applications on resource-constrained devices. On the other hand, the CNN and ResNet50 models shown to withstand even greater levels of compression highlighting their robustness and suitability for scenarios that demand maximum classification accuracy, even under stringent storage constraints. Importantly, we managed to run our optimized MobileNetV2 on an NVIDIA Jetson TX2 Developer Board, which proves the efficacy of deep learning-based diagnostics on edge devices, thus increasing the portability and usability of such tools in clinical settings. These findings give evidence of the need for a compression level and diagnostic performance balance in AI tools meant for healthcare. In future, we aim to assess newer lightweight architectures alongside developing diagnostic-driven adaptive compression strategies to extend the scalability of deep learning in medical imaging.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Neena K A: Conceptualized and designed the study, conducted the experiments, analyzed the data, and drafted the manuscript. Anil Kumar M N: Supervised the research work, provided critical insights during the study, and contributed to the manuscript’s review and editing. Both authors have read and approved the final version of the manuscript.

REFERENCES

[1] W. R. Hendee and E. R. Ritenour, *Medical Imaging Physics*, 4th ed. New York, USA: John Wiley & Sons. 2003.

[2] J. T. Norweck *et al.*, "ACR-AAPM-SIIM technical standard for electronic practice of medical imaging," *Journal of Digital Imaging*, vol. 26, no. 1, pp. 38–52, 2013. https://doi.org/10.1007/s10278-012-9522-2

[3] F. Liu, M. Hernandez-Cabronero, V. Sanchez *et al.*, "The current role of image compression standards in medical imaging," *Information*, vol. 8, no. 4, 2017. doi: 10.3390/info8040131

[4] A. Khashman and K. Dimililer, "Medical radiographs compression using neural networks and haar wavelet," in *Prof. IEEE EUROCON 2009*, 2009, pp. 1448–1453. doi: 10.1109/EURCON.2009.5167831

[5] M. Masmoudi, B. Jarboui, and P. Siarry, *Artificial Intelligence and Data Mining in Healthcare*, Switzerland: Springer, 2021.

[6] C. Yang, Y. Zhao, and S. Wang, "Deep image compression in the wavelet transform domain based on high frequency sub-band prediction," *IEEE Access*, vol. 7, pp. 52484–52497, 2019. doi: 10.1109/ACCESS.2019.2911403

[7] I. A. Urbaniak, "Using compressed JPEG and JPEG2000 medical images in deep learning: A review," *Applied Sciences*, vol. 14, no. 22, 10524, 2024. doi: 10.3390/app142210524

[8] National Electrical Manufacturers Association. (2019). The DICOM Standard. [Online]. Available: https://www.dicomstandard.org/current/

[9] O. S. Pianykh, *Digital Image Quality in Medicine*, Switherland: Springer, 2014. https://doi.org/10.1007/978-3-319-01760-0

[10] European Society of Radiology (ESR) comminucations@ myESR. org, "Usability of irreversible image compression in radiological imaging. A position paper by the European Society of Radiology (ESR)," *Insights into Imaging*, vol. 2, no. 2, pp. 103–115, 2011. https://doi.org/10.1007/s13244-011-0071-x

[11] Y. Y. Jo, Y. S. Choi, H. W. Park *et al.*, "Impact of image compression on deep learning-based mammogram classification," *Sci. Rep.*, vol. 11, no. 1, 7924, 2021. https://doi.org/10.1038/s41598-021-86726-w

[12] F. G. Zanjani, S. Zinger, B. Piepers *et al.*, "Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images," *J. Med. Imaging*, vol. 6, no. 2, 2019. doi: 10.1117/1.JMI.6.2.027501

[13] S. Krishnapriya and Y. Karuna, "Pre-trained deep learning models for brain MRI image classification", *Front. Hum. Neurosci.*, vol. 17, 2023. https://doi.org/10.3389/fnhum.2023.1150120

[14] M. Z. Khaliki and M. S. Başarslan, "Brain tumor detection from images and comparison with transfer learning methods and 3-layer CNN," *Sci Rep.*, vol. 14, no. 1, 2024. doi: 10.1038/s41598-024-52823-9

[15] B. B. Vimala, S. Srinivasan, S. K. Mathivanan *et al.*, "Detection and classification of brain tumor using hybrid deep learning models," *Sci Rep*, vol. 13, no. 1, 23029, 2023. https://doi.org/10.1038/s41598-023-50505-6

[16] Q. Min, X. Wang, B. Huang *et al.*, "Lossless medical image compression based on anatomical information and deep neural networks," *Biomedical Signal Processing and Control*, vol. 74, 103499, 2022. https://doi.org/10.1016/j.bspc.2022.103499

[17] P. K. Mallick, S. H. Ryu *et al.*, "Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network," *IEEE Access*, vol. 7, pp. 46278–46287, 2019. doi 10.1109/ACCESS.2019.2902252

[18] L. D. Chamain *et al*, "Quannet: Joint image compression and classification over channels with limited bandwidth," in *Proc. 2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 338–343. doi: 10.1109/ICME.2019.00066

[19] K. Dimililer, "DCT-based medical image compression using machine learning," *Signal, Image Video Process.*, vol. 16, no. 1, pp. 55–62, 2022. doi: 10.1007/s11760-021-01951-0

[20] B. Wang and J. Saniie, "Massive ultrasonic data compression using wavelet packet transformation optimized by convolutional autoencoders," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 3, pp. 1395–1405, 2021. doi: 10.1109/TNNLS.2021.3105367

[21] S. R. Sabbavarapu *et al.*, "A discrete wavelet transform and recurrent neural network based medical image compression for MRI and CT images," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 6, pp. 6333–6345, 2021. doi: 10.1007/s12652-020-02212-7

[22] Z. Liu *et al.*, "Machine vision guided 3D medical image compression for efficient transmission and accurate segmentation in the clouds," in *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12679–12688. doi: 10.1109/CVPR.2019.01297

[23] R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," *Procedia Computer Science*, vol. 197, pp. 198–207, 2022. doi: 10.1016/j.procs.2021.12.132

[24] Y. Chandola, J. Virmani, H. S. Bhadauria, and P. Kumar, "Lightweight end-to-end pre-trained CNN-based computer-aided classification system design for chest radiographs," in *Primers in Biomedical Imaging Devices and Systems*, Netherlands: Elsevier, 2021. https://doi.org/10.1016/B978-0-323-90184-0.00001-1

[25] C. Gunasundari *et al.*, "A novel approach for the detection of brain tumor and its classification using wavelet transform and support vector machine," *Scientific Reports*, vol. 15, no. 1, 2025. https://doi.org/10.1038/s41598-025-87934-4

[26] O. Abda and H. Naimi, "Enhanced brain tumor MRI classification using stationary wavelet transform, ResNet50V2, and LSTM networks," *ITEGAM-JETIA*, vol. 11, no. 51, pp. 127–133, 2025. https://doi.org/10.5935/jetia.v11i51.1457

[27] E. H. Yang, H. Amer, and Y. Jiang, "Compression helps deep learning in image classification," *Entropy*, vol. 23, no. 7, 881, 2021. https://doi.org/10.3390/e23070881

[28] S. A. Opee *et al.*, "ELW-CNN: An extremely lightweight convolutional neural network for enhancing interoperability in colon and lung cancer identification using explainable AI," *Healthcare Technology Letters*, vol. 12, no. 1, 2025. https://doi.org/10.1049/htl2.12122

[29] H. Zhou *et al.*, "Efficient human activity recognition on edge devices using DeepConv LSTM architectures," *Sci. Rep.*, vol. 15, no. 1, 13830, 2025. https://doi.org/10.1038/s41598-025-98571-2

[30] M. Nichparvar. (2022). Brain tumor MRI dataset. *Kaggle.* [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[31] K. M. Abubeker and S. Baskar, "B2-Net: An artificial intelligence powered machine learning framework for the classification of pneumonia in chest X-ray images," *Machine Learning: Science and Technology*, vol. 4, no. 1, 2023. Doi: 10.1088/2632-2153/acc30f

[32] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG 2000 still image coding system: An overview," *IEEE Trans. Consum. Electron.*, vol. 46, no. 4, pp. 1103–1127, 2000.

[33] C. Han *et al.*, "Toward variable-rate generative compression by reducing the channel redundancy," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 30, no. 7, pp. 1789–1802, 2020.

[34] D. K. Smith. (2012). Glymur: A Python interface for the JPEG 2000 reference software. [Online]. Available: https://glymur.readthedocs.io

[35] O. N. Belaid and M. Loudini, "Classification of brain tumor by combination of pre-trained VGG16 CNN," *J. Inform. Technol. Manage.*, vol. 12, pp. 13–25, 2020. doi: 10.22059/JITM.2020.75788

[36] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90

[37] M. Sandler *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510-4520.

[38] K. Memon *et al.*, "Edge computing for AI-based brain MRI applications: A critical evaluation of real-time classification and segmentation," *Sensors*, vol. 24, no. 21, 2024. doi: 10.3390/s24217091

[39] K. Dong *et al.*, "MobileNetV2 model for image classification," in *Proc. 2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, 2020, 476–480. doi: 10.1109/ITCA52113.2020.00106

[40] S. Ram *et al.*, "Leveraging diverse CNN architectures for medical image captioning: DenseNet-121, MobileNetV2, and ResNet-50 in ImageCLEF 2024," in *Proc. CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org*, 2024.

[41] M. H. Ferris *et al.*, "Using ROC curves and AUC to evaluate performance of no-reference image fusion metrics," in *Proc. 2015 National Aerospace and Electronics Conference (NAECON)*, 2015, pp. 27–34. doi: 10.1109/NAECON.2015.7443034