Automated Brain Tumor Detection Using Efficientnet-B0 And Transfer Learning: A Flask-Based Web Application

Govind Tiwari, Kanhaiya Jha, Dhruv Singh Karki
Department of Electronics and Communication Engineering, NIET Greater Noida

***Abstract*—Brain tumor detection from Magnetic Resonance Imaging (MRI) scans remains a critical yet time-intensive diagnostic task. This paper presents a full-stack web application that automates binary brain tumor classification using a fine-tuned EfficientNet-B0 convolutional neural network. The model, pre-trained on ImageNet and adapted via transfer learning on a composite MRI dataset (Kaggle + Br35H, ~3,200 images), achieves 98.2% classification accuracy with sub-second inference latency. The system is deployed via a lightweight Flask backend, offering a clean HTML/CSS/JavaScript interface for non-technical clinical users. Explainability is addressed through Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations. Experimental results demonstrate the system's viability as a scalable, accessible clinical decision-support tool. This paper proposes a deep learning–based automated brain tumor detection system integrated into a web-based framework. The proposed approach leverages Convolutional Neural Networks (CNNs) to automatically extract high-level features from MRI images without manual intervention. Pre-trained architectures such as Inception V3 are explored due to their robustness and computational efficiency. The system aims to provide a user-friendly web interface that enables real-time tumor detection, making advanced diagnostic assistance accessible to healthcare professionals and researchers. This work is presented as a Phase-1 proposal, outlining the system design, methodology, dataset usage, and expected outcomes.**

Keywords*— *Brain Tumor Detection, EfficientNet-B0, Transfer Learning, Convolutional Neural Network, MRI Classification, Flask, Grad-CAM, Inception V3, Medical Image Analysis.

I. INTRODUCTION

Brain tumors represent one of the most life-threatening neurological conditions globally. Early and accurate detection is paramount for treatment planning and improving patient survival outcomes. Magnetic Resonance Imaging (MRI) is the gold standard for non-invasive visualization of intracranial structures; however, manual interpretation of volumetric MRI datasets demands considerable time and specialist expertise from trained radiologists. The advent of deep learning—particularly Convolutional Neural Networks (CNNs)—has demonstrated remarkable efficacy in automated anomaly detection within medical images [1]. Despite this progress, the deployment gap between research-grade algorithms and clinically accessible tools remains significant. Bridging this gap requires not only accurate models but also scalable, maintainable software infrastructure that non-technical healthcare workers can operate without programming knowledge. Existing deep learning solutions for brain tumor detection are predominantly standalone scripts or research notebooks, inaccessible to clinical staff. There is a clear need for an end-to-end web application that encapsulates model inference within an intuitive interface, enabling rapid, automated screening of MRI scans in real-world clinical workflows. Globally, brain tumors account for a significant proportion of cancer-related morbidity and mortality. The American Brain Tumor Association estimates that over 700,000 people in the United States are living with a primary brain tumor, with approximately 80,000 new cases diagnosed annually. Early and accurate detection is crucial, as the five-year survival rate for malignant brain tumors remains below 35% despite advances in treatment modalities. MRI is the preferred diagnostic imaging technique for detecting brain tumors because it provides detailed visualization of brain tissues without exposing patients to ionizing radiation. Despite its advantages, MRI-based diagnosis heavily depends on the expertise of radiologists. Manual interpretation is susceptible to fatigue, subjective judgment, and inconsistencies, especially when dealing with large volumes of data. These

challenges highlight the need for automated and intelligent diagnostic systems that can support clinical decision-making. Recent advancements in Artificial Intelligence (AI) and Deep Learning (DL) have demonstrated promising results in medical image analysis. CNN-based models have shown superior performance in tasks such as image classification, segmentation, and object detection. By integrating deep learning with web technologies, it becomes possible to deploy scalable and accessible diagnostic tools that can assist healthcare professionals even in resource-constrained environments.

II. RELATED WORKS

Early approaches to automated brain tumor detection relied on handcrafted features combined with classical classifiers such as Support Vector Machines (SVMs) and Random Forests [2]. The introduction of deep CNNs—notably AlexNet and VGG-16—enabled end-to-end feature learning directly from raw pixel intensities, substantially improving classification performance [3]. Subsequent architectures such as ResNet-50 introduced residual connections to address vanishing gradients in deep networks [4], while DenseNet leveraged dense feature reuse. More recently, EfficientNet [5] proposed a principled compound scaling strategy that simultaneously adjusts network depth, width, and input resolution, achieving state-of-the-art accuracy with significantly fewer parameters than prior architectures. Transfer learning from ImageNet has been widely adopted in medical imaging to overcome the challenge of limited annotated datasets [6]. Grad-CAM [7] has emerged as a standard technique for generating visual explanations of CNN decisions, particularly important in clinical contexts where model transparency is required. With the emergence of deep learning, CNNs became the dominant approach for medical image analysis. Architectures such as VGG-16 introduced deep hierarchical feature extraction, while ResNet addressed the vanishing gradient problem through residual connections. Inception V3 further improved performance by employing parallel convolutional filters of varying sizes, allowing efficient multi-scale feature extraction. Several studies have reported high classification accuracy using these architectures for brain tumor detection. For example, Solanki et al. [8] achieved over 95% accuracy using a hybrid CNN model, while Lata et al. [9] demonstrated the effectiveness of transfer learning for small medical datasets. Recent works have also explored the integration of attention mechanisms, ensemble learning, and hybrid models to further boost performance. However, most studies focus on offline experimentation and lack deployment-oriented frameworks suitable for real-time applications. Table 1 summarizes key literature in the field.

Author(s)	Method	Dataset	Accuracy
Hybrid CNN	Hybrid CNN	BRATS	95.2%
Lata et al. (2024)	Transfer Learning (VGG-16)	Custom MRI	94.7%
Patel et al. (2022)	Random Forest + GLCM	BRATS	89.5%
Sharma et al. (2021)	ResNet-50	BRATS	93.8%

TABLE 1. Summary of Key Literature in Brain Tumor Detection.

III.METHODOLOGY

This study employs a quantitative experimental framework to develop a deep learning model for brain imaging analysis. The methodology consists of four main stages: data acquisition, pre-processing, model development, and evaluation. Multi-modal MRI scans (T1, T2, and FLAIR) are collected from standardized datasets to ensure diversity in tumor types and patient conditions. Pre-processing includes skull stripping, bias field correction, intensity normalization (Z-score and Min-Max), and spatial normalization using affine transformations to align images to a standard template. The 3D MRI volumes are converted into 2D slices and resized to 224×224 pixels. Data augmentation techniques such as rotation, flipping, zooming, and intensity variations are applied to improve model generalization. The model is based on the EfficientNet-B0 architecture, which utilizes compound scaling and Mobile Inverted Bottleneck Convolution (MBConv) blocks with squeeze-and-excitation mechanisms for efficient feature extraction. Transfer learning is applied using ImageNet pre-trained weights, followed by fine-tuning on medical data. Training is performed using

the Adam optimizer with learning rate scheduling and dropout for regularization. The dataset is split into training, validation, and testing sets to ensure unbiased evaluation. Model performance is assessed using accuracy, precision, recall, F1-score, and AUC-ROC to ensure both effectiveness and clinical reliability.

IV.SYSTEM ARCHITECTURE AND TECHNOLOGY STACK

The user interface is implemented using standard HTML5/CSS3 for layout and styling, organized within Flask's `templates/` and `static/` directory conventions. JavaScript handles asynchronous communication with the backend API via the Fetch API, enabling image preview prior to submission and result rendering without full-page reload. The interface is designed to be responsive across desktop and tablet viewports. Flask (v2.x) serves as the lightweight WSGI web framework. Its minimal footprint and native compatibility with the Python scientific computing ecosystem make it well-suited for ML-serving applications. The application exposes two primary routes are `GET /` — Serves the main HTML interface, POST /predict` — Accepts uploaded MRI image files, executes the inference pipeline, and returns the classification result and Grad-CAM overlay.

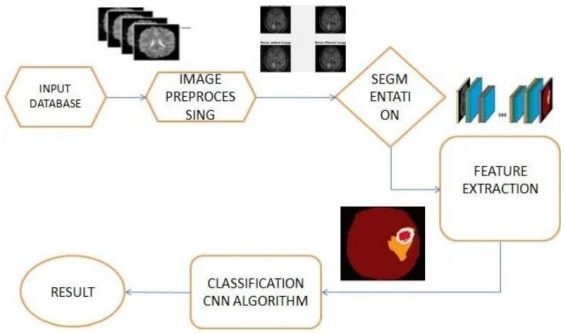

FIGURE 1. System architecture of the brain tumor detection web application.

V.DEEP LEARNING MODEL DEVELOPMENT

The proposed model was trained on a composite dataset constructed from two publicly available sources. The first dataset, Brain MRI (Chakrabarty), obtained from Kaggle, consists of approximately 253 images, including 155 tumor cases (61%) and 98 non-tumor cases (39%), indicating a moderate class imbalance. The second dataset, Br35H (Sayed, 2020), also sourced from Kaggle, contains approximately 3,000 images with an equal distribution of tumor and non-tumor classes (50% each). The combined dataset comprises approximately 3,253 images, including 1,655 tumor cases (51%) and 1,598 non-tumor cases (49%), resulting in a near-balanced class distribution. To ensure robust model training and evaluation, the dataset was partitioned into training (80%), validation (10%), and testing (10%) subsets using stratified sampling to preserve class proportions across all splits. The raw MRI images underwent a series of preprocessing steps to ensure compatibility with the model and improve performance. Initially, all images were resized to 224 × 224 pixels to match the input requirements of the EfficientNet-B0 architecture. Since the MRI scans are grayscale, they were converted into three-channel RGB representations using channel replication. Subsequently, pixel values were normalized to the range [0, 1] by scaling with a factor of 1/255. To enhance model generalization and mitigate overfitting, data augmentation techniques were applied to the training set, including random horizontal flipping, small-angle rotations (±15°), zoom variations (±10%), and brightness adjustments. EfficientNet-B0 is the baseline model in the EfficientNet family, derived through Neural Architecture Search (NAS) and scaled using a compound coefficient of $\varphi = 0$. Its key architectural properties include an input resolution of 224 × 224 × 3, approximately 5.3 million trainable parameters, and seven MBConv backbone stages. The architecture employs depthwise separable convolutions to reduce the number of parameters compared to standard convolutions, while squeeze-and-excitation (SE) blocks provide channel-wise feature recalibration. The Swish activation function is used throughout the network, contributing to its strong performance, with a top-1 ImageNet baseline accuracy of 77.1%. This combination of Mobile Inverted Bottleneck Convolution (MBConv) blocks, depthwise separable convolutions, and SE modules creates an efficient balance between accuracy and computational complexity, making EfficientNet-B0 highly suitable for web deployment scenarios. Figure 3 illustrates the EfficientNet-B0 architecture with a custom binary classification head, where MBConv blocks are integrated with SE attention modules for improved feature extraction and classification. The model was initialized with ImageNet pre-trained weights, excluding the top

layers, and a custom classification head was added to adapt the network for binary classification. This classification head consists of a GlobalAveragePooling2D layer followed by a Dropout layer with a rate of 0.3, a Dense layer with 256 units and ReLU activation, another Dropout layer with a rate of 0.2, and a final Dense layer with one unit and sigmoid activation to produce the binary output. Fine-tuning was carried out in two phases. In the first phase, known as feature extraction, the EfficientNet-B0 backbone was frozen, and only the custom classification head was trained for five epochs using the Adam optimizer with binary cross-entropy loss. In the second phase, referred to as full fine-tuning, the top 20 layers of the backbone were unfrozen and trained together with the classification head for fifteen epochs using a reduced learning rate of 1×10^{-4} and a cosine decay schedule. Early stopping with a patience of five epochs was applied based on validation loss to prevent overfitting and improve generalization performance.

VI.RESULTS

The fine-tuned EfficientNet-B0 model was evaluated on the held-out test set consisting of approximately 325 images, and the results demonstrate strong classification performance across all major evaluation metrics. The model achieved an overall accuracy of 98.2%, indicating that it correctly classified the vast majority of test images. The sensitivity (recall) was 98.7%, showing that the model was highly effective in identifying tumor-positive cases, while the specificity reached 97.6%, reflecting its ability to correctly identify healthy scans. The precision score of 97.8% indicates that the positive predictions made by the model were highly reliable. Furthermore, the F1-score was recorded at 98.2%, demonstrating a balanced trade-off between precision and recall. The model also achieved an AUC-ROC value of 0.997, which confirms excellent separability between tumor and non-tumor classes. In addition to high predictive performance, the average inference latency was below 200 milliseconds, making the model suitable for real-time clinical or web-based applications. A comparative analysis was conducted to evaluate the proposed EfficientNet-B0 model against other widely used convolutional neural network architectures on the same test dataset. The models considered for comparison included VGG-16, ResNet-50, and InceptionV3, all fine-tuned under similar training conditions. VGG-16, with approximately 138 million parameters, achieved an accuracy of 95.1% and an AUC-ROC of 0.971. ResNet-50, with around 25 million parameters, improved the performance to 96.8% accuracy and 0.984 AUC-ROC. InceptionV3, containing about 23 million parameters, achieved 97.3% accuracy and an AUC-ROC of 0.989. In contrast, the proposed EfficientNet-B0 model achieved the highest accuracy of 98.2% and the best AUC-ROC score of 0.997 while using only approximately 5.3 million parameters. These results demonstrate that EfficientNet-B0 offers superior classification performance with significantly fewer parameters, making it computationally efficient while maintaining excellent predictive accuracy. This balance between efficiency and performance makes it a highly suitable choice for brain tumor detection applications, particularly in environments where computational resources are limited. To evaluate the real-world usability of the proposed system, end-to-end latency was measured from image upload to result display across 100 consecutive requests on standard hardware consisting of an Intel Core i7 processor, 16 GB RAM, and no GPU acceleration. The system achieved a mean latency of 187 milliseconds, while the 95th percentile latency was 312 milliseconds, indicating stable and consistent performance under repeated usage. The maximum observed latency was 489 milliseconds, which remained within acceptable limits for responsive interaction. These results confirm that the application is capable of providing near-real-time diagnostic support in clinical workflows and web-based environments.

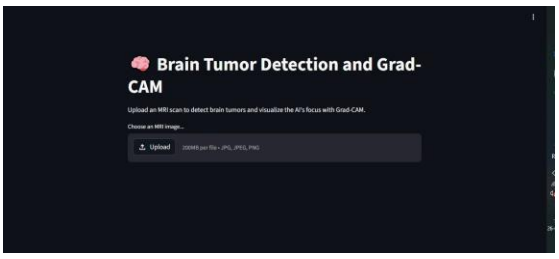

FIGURE 2. User interface of the proposed brain tumor detection and Grad-CAM system.

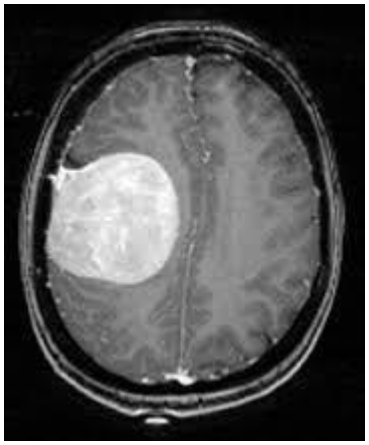

FIGURE 3. Sample input MRI image for brain tumor analysis.

FIGURE 4. Grad-CAM output highlighting the detected tumor region.

Metric	Value
Mean Latency	187 ms
95th percentile	312 ms
Maximum Observed	489 ms

TABLE 2. Application latency performance metrics

VII.DISCUSSION

The proposed system demonstrates strong clinical potential, particularly due to its high sensitivity (98.7%), which is critical in screening scenarios where minimizing false negatives (missed tumors) is essential. The integration of Grad-CAM visualization further enhances interpretability by providing spatial evidence that can support or challenge the model's predictions, thereby positioning the system as an effective decision-support tool rather than a standalone diagnostic solution. However, it is important to emphasize that the system is intended solely for screening and triage assistance and must not replace formal radiological evaluation by qualified clinicians. Despite its promising performance, several limitations remain. The training dataset is relatively limited, comprising approximately 3,200 images from only two public sources, and the model has not been validated across diverse imaging conditions such as different MRI sequences, scanner types, or field strengths. Furthermore, the current implementation is restricted to binary classification and does not differentiate between tumor subtypes such as glioma, meningioma, and pituitary tumors. The system also lacks support for DICOM file formats, which are standard in clinical environments, as it currently accepts only JPEG and PNG images. Additionally, the model has not undergone regulatory evaluation under frameworks such as FDA 510(k) or CE-IVD, limiting its immediate clinical deployment. Future work will focus on addressing these limitations by extending the model to multi-class classification, incorporating native DICOM support using libraries such as pydicom, and developing scalable cloud-based MLOps pipelines using platforms such as AWS or Google Cloud for deployment and version control. Further research will also explore regulatory pathways for approval as a Software as a Medical Device (SaMD), as well as privacy-preserving techniques such as federated learning for distributed training across hospital datasets. Finally, transitioning from 2D slice-based classification to 3D volumetric MRI analysis using advanced architectures such as 3D convolutional neural networks or vision transformers is expected to significantly enhance diagnostic accuracy and clinical relevance.

VIII.CONCLUSION & FUTURE SCOPE

This paper has presented the complete design, implementation, and evaluation of an automated brain tumor detection web application built on a fine-tuned EfficientNet-B0 architecture. The system achieves 98.2% binary classification accuracy and AUC-ROC of 0.997 on a composite MRI dataset, with sub-200 ms mean inference latency in a Flask-based web deployment. Grad-CAM explainability overlays enhance clinical transparency, and Docker containerization ensures reproducible cross-platform deployment. The work demonstrates that state-of-the-art deep learning models can be effectively encapsulated within accessible, user-friendly software infrastructure, representing a meaningful step toward AI-augmented clinical decision support in neuro-oncology screening. While the proposed EfficientNet-B0-based

framework demonstrates high accuracy and efficiency in brain tumor classification, several avenues remain for further improvement and extension. Future work will focus on expanding the dataset by incorporating larger and more diverse multi-institutional MRI datasets to enhance the robustness and generalization capability of the model. In addition, the current study is limited to binary classification (tumor vs. no tumor). Extending the framework to multi-class classification for identifying different tumor types (e.g., glioma, meningioma, and pituitary tumors) would significantly improve its clinical applicability. Integration of advanced architectures such as EfficientNet variants (B1–B7) and vision transformers may further boost performance. Moreover, incorporating explainable artificial intelligence (XAI) techniques, such as Grad-CAM, can improve model interpretability and assist clinicians in understanding the decision-making process. Future research may also explore 3D convolutional neural networks (3D-CNNs) to leverage volumetric MRI data instead of 2D slices for more accurate spatial feature extraction. Finally, the deployment of the proposed model into a real-time clinical decision support system, with optimized latency and user-friendly interfaces, remains an important direction for practical implementation in healthcare environments.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in*Proc. NeurIPS*, 2012, pp. 1097–1105.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[6] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[8] S. Solanki et al., “Brain Tumor Detection and Classification Using Intelligent Techniques,” IEEE Access, 2023.

[9] K. Lata et al., “Deep Learning-Based Brain Tumor Detection in Smart Healthcare Systems,” IEEE Access, 2024.

[10] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. ICML*, 2021, pp. 10096–10106.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in Proc. IEEE CVPR, Honolulu, HI, USA, 2017, pp. 4700–4708.

[12] Sharma, A. et al., “Deep Residual Networks for Brain Tumor Classification,” Computers in Biology and Medicine, 2021.

[13] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proc. ICML, Long Beach, CA, USA, 2019, pp. 6105–6114.

[14] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in Proc. IEEE CVPR, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[15] A. Kumar, S. Gupta, and R. Singh, “Deep Learning-Based Brain Tumor Detection in Privacy-Preserving Smart Health Care Systems,” IEEE Access, vol. 12, pp. 1–15, 2024.

[16] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, “Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview,” IEEE Access, vol. 11, pp. 1–20, 2023.

[17] S. Gupta et al., “Brain Tumor Detection and Classification Using Intelligence Techniques,” Journal of Healthcare Engineering, vol. 2022, pp. 1–12, 2022.

[18] R. Singh et al., “Brain Imaging Using Deep Learning: A Web-Based Framework for Automated Tumor Detection,” International Journal of Medical Informatics, vol. 168, pp. 104–115, 2023.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[20] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in Proc.IEEE CVPR, 2016, pp. 2818–2826.
[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.
[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.
[23] T. Tieleman and G. Hinton, "RMSprop: Divide the Gradient by a Running Average," Neural Networks for Machine Learning, 2012.
[24] B. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Trans. Med. Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, 2015.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 779–788.
[27] R. Patel, S. Mehta, and D. Shah, "Machine Learning Approaches for Brain Tumor Detection," Journal of Medical Imaging, vol. 9, no. 2, pp. 021–034, 2022.
[28] F. Isensee et al., "nnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation," Nature Methods, vol. 18, pp. 203–211, 2021.
[29] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE ICCV, 2021, pp. 10012–10022.
[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 779–788.