Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3518~3528
e-ISSN: 2548-6861

Efficient Feature Extraction Using MobileNetV2 and EfficientNetB0 for Multi-Class Brain Tumor Classification

Hemas Anggita Amelia [1]*, Majid Rahardi [2]*
*Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta
anggitaamelia@students.amikom.ac.id [1] majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-09-25
Revised 2025-11-13
Accepted 2025-11-15

Keyword:

EfficientNetB0,
Machine Learning,
Lightweight CNN,
Brain Tumor Classification.

ABSTRACT

Brain tumor classification in MRI is complicated by the similarity of imaging features across multiple tumor classes. This study evaluates the use of lightweight convolutional neural network (CNN) architectures as feature extractors combined with machine learning classifiers for multi-class classification. MobileNetV2 and EfficientNetB0 were used to extract fixed-length feature representations, which were then classified using Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors. The evaluation used stratified five-fold cross-validation, and performance was measured with accuracy, F1-score, and Matthews Correlation Coefficient (MCC). Results show that EfficientNetB0 features paired with SVM achieved the highest test accuracy (98.5%), while Logistic Regression also yielded competitive performance (97.1%). Class-wise analysis indicated strong results for pituitary and non-tumor cases. This work shows that lightweight CNN-based feature extraction may serve as a practical direction for improving multi-class brain tumor MRI classification, with potential benefits for applications in resource-limited environments.

I. INTRODUCTION

Brain tumors are among the most challenging conditions in neurology, with a considerable impact on public health. Recent estimates report over 320,000 new cases of brain and central nervous system cancers and nearly 250,000 related deaths each year [1]. The Global Burden of Disease (GBD) study further indicates that the incidence of brain tumors continues to rise across various age groups [2]. These statistics emphasize the importance of reliable diagnostic support systems to assist in tumor recognition and classification.

Magnetic Resonance Imaging (MRI) is the standard technique for detecting and evaluating brain tumors because of its high-resolution images and superior soft-tissue contrast [3]. Despite these advantages, interpretation of MRI scans remains difficult. Radiologists often face challenges in differentiating between tumor subtypes with overlapping visual patterns, such as gliomas and meningiomas, which can lead to variations in diagnosis [4]. This difficulty, combined with the time required for expert analysis, motivates the search for automated solutions that can assist in classification tasks.

Deep convolutional neural networks (CNNs) have advanced medical image analysis significantly. Models such as ResNet, DenseNet, and EfficientNet have shown strong performance in brain tumor classification by automatically learning rich feature representations [5], [6]. However, end-to-end CNN approaches typically demand large labeled datasets and high computational resources, and they often provide limited interpretability [7]. These limitations reduce their practicality, especially in environments with restricted access to large-scale data or advanced hardware.

Hybrid frameworks provide a potential alternative. Recent studies highlight the effectiveness of pretrained CNNs as feature extractors combined with traditional machine learning classifiers [8]. For instance, Deepak et al. used CNN-based feature extraction with SVM to reach a 95.82% classification accuracy [9]. Moreover, lightweight CNN architectures like MobileNetV2 and EfficientNetB0 have gained attention due

to their compact model size and high efficiency, making them ideal for use in resource-constrained environment [10]

Despite these advancements, gaps in methodology remain. Yagis et al. [11] cautioned that improper preprocessing and data leakage may artificially inflate model performance, while Benaouali et al. [12] emphasized the importance of rigorous validation when using CNN features for medical image analysis. Moreover, many studies report only single train/test split results without applying cross-validation or statistical testing, limiting the reliability and generalizability of their findings. In addition, Gómez-Guzmán et al. [13] noted that high-performance computing resources are often required to train end-to-end CNN models, which can limit their applicability in settings with limited computational capacity.

To address these gaps, this study conducts a systematic comparison of two lightweight convolutional neural networks such as MobileNetV2 and EfficientNetB0 as fixed feature extractors for multiclass brain tumor MRI classification. Both models were evaluated under identical experimental conditions, including consistent preprocessing, classifier configurations, and leakage-free stratified cross-validation. The objective of this work is not to propose a new architecture but to assess how effectively compact CNN backbones can transfer learned representations to the task of distinguishing gliomas, meningiomas, pituitary tumors, and no-tumor cases. By incorporating statistical validation and per-class performance analysis, this study aims to explore and evaluate different lightweight convolutional neural networks (CNNs) such as MobileNetV2 and EfficientNetB0, to determine which model performs better under resource constraints. The research also applies preprocessing, cross-validation, and statistical testing as part of its evaluation process, aiming to provide clarity and consistency in the analysis.

II. METHOD

This study presents a comparative analysis of feature representation between MobileNetV2 and EfficientNetB0 for multi-class brain tumor MRI classification. Both models are employed as pretrained feature extractors, while machine learning classifiers are used in the final classification stage. This hybrid design was chosen to reduce training complexity and provide a fair assessment of how well each backbone captures discriminative features from MRI images. End-to-end CNN training was not pursued due to the limited dataset size and the intention to focus on efficient feature extraction strategies suitable for constrained hardware environments.

To ensure methodological rigor, the experimental pipeline was designed with separation between training and testing data. All preprocessing, normalization, and feature extraction steps were performed independently within each fold of cross-validation, ensuring that no information from the test data influenced model training. This design minimizes bias and allows an unbiased comparison of the representational quality of the two CNN architectures.

The overall workflow of the study is illustrated in Figure 1, which outlines the steps from dataset preparation to final evaluation. After preprocessing, deep features were extracted from MobileNetV2 and EfficientNetB0, then passed into machine learning classifiers including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Model performance was assessed through stratified five-fold cross-validation, followed by evaluation on test set. Finally, statistical validation was applied to compare classifiers across backbones, ensuring that observed performance differences were robust and not due to random variation. The following subsections describe each stage of this workflow in detail.

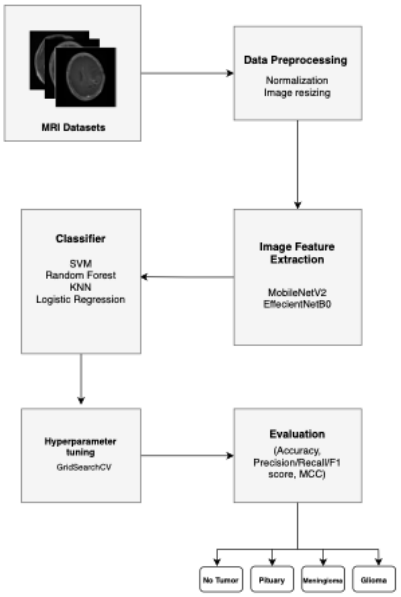

Figure 1. Framework diagram of the proposed methodology.

A. Datasets

This study utilizes a publicly available brain tumor MRI dataset obtained from Kaggle [14]. The dataset aggregates MRI scans from three established public sources: the Br35H dataset, the SARTAJ dataset, and a brain tumor collection from Figshare. This combined dataset contains four diagnostic categories: glioma, meningioma, pituitary tumor, and no tumor, providing a comprehensive benchmark for multi-class brain tumor classification.

The dataset consists of 7,023 axial T1-weighted contrast-enhanced MRI scans with varying original resolutions. The relatively balanced class distribution in Table 1 which minimizes the risk of classification bias during model training and evaluation.

TABLE 1
DISTRIBUTION OF MRI IMAGES ACROSS TUMOR CLASSES

Class	Total
Glioma	1621
Meningioma	1645
Pituitary	1757
No Tumor	2000
Total	**7023**

A key characteristic noted by the dataset curator is that the original images have varying sizes and may contain extra margins. As explicitly mentioned in the dataset documentation, resizing the images after pre-processing is recommended to improve model accuracy [14]. This informed our pre-processing strategy, detailed in the following section, to standardize the input dimensions for the deep learning models, while Figure 2 presents representative examples of MRI scans for each category. This balanced organization across classes makes the dataset suitable for benchmarking both deep learning and machine learning approaches.

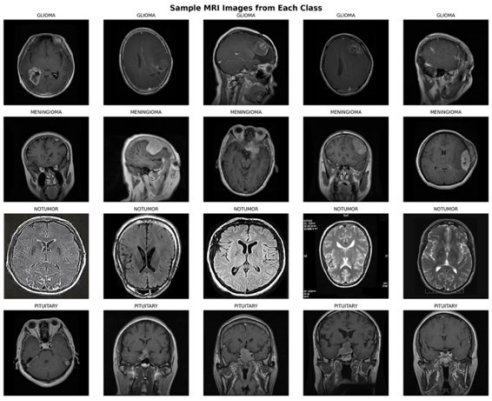

Figure 2. Sample images from MRI dataset.

B. Preprocessing

The dataset underwent comprehensive preprocessing to ensure methodological integrity and compatibility with deep learning architectures. A critical sanity check confirmed no filename overlap between the training (5,712 images) and testing (1,311 images) subsets, verifying the absence of patient data leakage between partitions. Preprocessing applied before data splitting can lead to data leakage, where information from the test set inadvertently influences the training process.

1) Image Standardization: All MRI images were uniformly resized to 224×224 pixels using bilinear interpolation to meet the input dimension requirements of MobileNetV2 and EfficientNetB0. Grayscale images were converted to 3-channel RGB format by duplicating the single channel across all three channels to satisfy CNN input requirements while preserving original intensity information..

2) Normalization: Pixel intensities were normalized using architecture-specific preprocessing functions. For MobileNetV2, images were scaled to the [-1, 1] range using the mobilenet_preprocess function, while EfficientNetB0 used the efficientnet_preprocess function which applies channel-wise normalization based on ImageNet statistics. This ensures compatibility with the pretrained weights from ImageNet.

While data augmentation is commonly employed in deep learning to improve generalization, it was deliberately excluded from this study to maintain consistent feature representations across the dataset. The use of static feature extraction without fine-tuning meant that augmented variations could introduce inconsistencies in the feature space, as the pretrained models were not adapted to these artificial transformations. This contamination artificially inflates performance metrics, creating an overly optimistic assessment of the model's capabilities and undermining the reliability and generalizability of the results [15]. The conservative approach establishes a clear baseline for evaluating the raw feature representation quality of the pretrained architectures, isolated from the effects of data augmentation.

Furthermore, to prevent data leakage and evaluation bias, all preprocessing operations were implemented with strict protocols where normalization parameters were calculated exclusively from the training portion of each cross-validation fold and subsequently applied to validation and test data, ensuring that no information from validation or test sets influenced the training process at any stage.

C. Feature Extraction

Both MobileNetV2 and EfficientNetB0 were employed as static feature extractors using weights pretrained on the ImageNet dataset. No fine-tuning or backpropagation was performed on the brain tumor dataset, preserving the generic feature representations learned from natural images. The classification layers of both architectures were removed, and global average pooling was applied to generate fixed-length 1,280-dimensional feature vectors from each MRI image. This static feature extraction approach allows for direct comparison of representational quality between architectures while maintaining computational efficiency and preventing overfitting on the medical imaging dataset.

The extraction stage is central to our comparative analysis, as it isolates the representational capabilities of different CNN backbones from classifier effects. Instead of fine-tuning the networks end-to-end, both MobileNetV2 and EfficientNetB0 were employed as fixed feature extractors. Their classification layers were removed, and the global average pooling output was used to generate fixed-length feature vectors. This design allows a direct evaluation of feature

representation quality while avoiding confounding factors introduced by retraining.

Each model produced feature vectors of different dimensionalities due to architectural differences:

1) *MobileNetV2*: Initialized with ImageNet-pretrained weights, producing a 1,280-dimensional feature vector per image. The architecture is based on depth wise separable convolutions arranged in inverted residual blocks with linear bottlenecks, which provide an efficient yet expressive representation [16]. Input images were scaled to following MobileNetV2's channel-wise normalization. Features were extracted in batches of 64, ensuring dropout was disabled and batch normalization layers operated consistently. The resulting feature matrices were saved separately for the training and test subsets, enabling reproducible downstream classification.

2) EfficientNetB0: Configured with parameters to produce 1,280-dimensional feature vectors, enabling a direct comparison with MobileNetV2. Unlike MobileNetV2, EfficientNetB0 employs compound scaling to balance depth, width, and resolution, resulting in more expressive features within a lightweight architecture [17]. Feature extraction was performed in batches of 64 to ensure deterministic behavior. Using identical feature dimensions and standardized processing steps allowed a fair comparison of representational quality between the two backbones without confounding artifacts.

MobileNetV2 and EfficientNetB0 were chosen as feature extractors because of their lightweight architecture and computational efficiency, which makes them attractive for use in medical imaging tasks under resource limitations. The smaller complexity of MobileNetV2 and EfficientNetB0 indicating that it may offer a favorable trade-off between accuracy and efficiency for future exploration in clinical decision support systems.

D. Classification

Accurate classification is a crucial stage in medical imaging pipelines, as it determines the ability of an automated system to distinguish between disease subtypes and provide reliable decision support [18]. In brain tumor analysis, this step translates extracted image features into meaningful categories. Since our study focuses on evaluating the representational power of CNN-derived features, the classification stage plays a central role in assessing how effectively these features separate tumor classes.

To capture a broad range of decision-making paradigms, four traditional classifiers were employed: Support Vector Machine (SVM), Logistic Regression, Random Forest (RF), and K-Nearest Neighbors (KNN). SVM with a radial basis function (RBF) kernel was selected for its ability to model non-linear class boundaries, where the kernel function is defined as

$$K(x_i, x_j) = \left(-\gamma \left|x_i - x_j\right|^2\right) \quad (1)$$

with γ controlling the influence of each training instance. Logistic Regression was implemented in its multinomial form, estimating class membership probabilities through the softmax function:

$$P(y = k \mid x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1} \exp\left(w_j^T x + b_j\right)} \quad (2)$$

where y is the target class and K is number of tumor classes. Random Forest, as an ensemble of decision trees, was included to improve robustness and reduce variance, while KNN provided a non-parametric baseline that directly reflects the geometry of the feature space.

Scaling operations were embedded where necessary, and the same feature vectors extracted from MobileNetV2 and EfficientNetB0 were provided to each classifier under identical conditions. This uniform design ensured that observed performance differences arose primarily from the representational capacity of the CNN backbones, rather than inconsistencies in classifier implementation, thereby allowing a fair and unbiased comparison.

E. Hyperparameter Tuning

Hyperparameter optimization was conducted to ensure fair and unbiased comparisons across classifiers [19]. For each algorithm, hyperparameters were tuned using stratified five-fold cross-validation applied only to the training set. This design preserved the independence of the held-out test set and prevented optimistic bias. Tuning was implemented through grid search, with candidate ranges defined based on prior studies and practical constraints.

For Support Vector Machine, the penalty parameter C and kernel coefficient γ were varied across predefined grids. Logistic Regression was tuned by adjusting the regularization strength C. Random Forest parameters included the number of trees ($n_{\text{estimators}}$) and maximum depth of trees. For K-Nearest Neighbors, the number of neighbors (k) was varied.

F. Evaluation

Model performance was assessed using a two-phase protocol. First, stratified 5-fold cross-validation was applied exclusively to the training set for hyperparameter tuning, ensuring balanced class representation and preventing data leakage. After selecting the best configuration, models were retrained on the full training set and evaluated on the test set to provide unbiased estimates of generalization capability.

A comprehensive set of metrics was used to capture both overall and class-specific performance. These included accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). Precision, recall, and F1 were reported in macro averaged forms to account for potential class imbalance. In addition, per-class precision, recall, and F1-scores were provided for glioma, meningioma, pituitary,

and no tumor categories to highlight specific strengths and weaknesses. The following are the performance metrics that will be utilized.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

accuracy measures the overall proportion of correct predictions.

$$\text{Precision}_k = \frac{TP_k}{TP_k+FP_k} \quad (4)$$

precision evaluates how many predicted positives for a class were actually correct.

$$\text{Recall}_k = \frac{TP_k}{TP_k+FN_k} \quad (5)$$

recall (Sensitivity) measures how many true instances of a class were successfully identified.

$$F1_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \quad (6)$$

f1-score is the harmonic mean of precision and recall, balancing both aspects.

For multiclass MCC, we adopted the generalized definition:

$$MCC = \frac{c \cdot s - \sum_k p_k \cdot t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \quad (7)$$

where c is the number of correctly classified samples, s is the total number of samples, p_k is the number of instances predicted for class k, and t_k is the number of true instances for class k.

MCC provides a correlation coefficient between predicted and true labels, ranging from –1 (total disagreement) to +1 (perfect prediction). Unlike accuracy, MCC accounts for all four outcomes (TP, TN, FP, FN) and is more robust in imbalanced multi-class problems.

III. Result and Discussion

This research presents comparative analysis of feature representation quality between MobileNetV2 and EfficientNetB0 architectures for multi-class brain tumor MRI classification. This section details our findings through quantitative performance metrics, feature space visualization, and per-class analysis on these two prominent CNN architectures for medical image feature extraction.

A. Feature Representation

To gain insights into the quality of feature representations extracted by MobileNetV2 and EfficientNetB0, we conducted t-SNE visualization of the 1,280-dimensional feature spaces. Figure 3 and Figure 4 presents the two-dimensional embeddings of training features for both architectures.

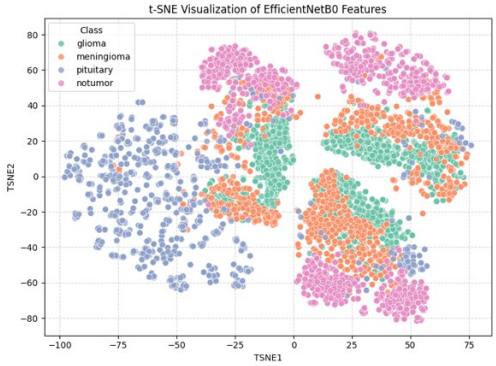

Figure 3. Feature of EfficientNetB0 represents a brain MRI sample colored by tumor type.

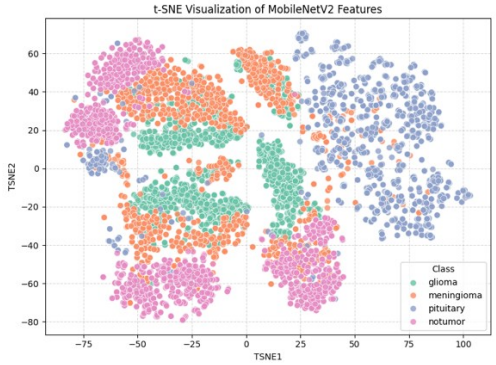

Figure 4. Feature of MobileNetV2 represents a brain MRI sample colored by tumor type.

The visualization reveals that EfficientNetB0 features form more distinct and well-separated clusters for all four classes (glioma, meningioma, pituitary tumor, and no tumor) compared to MobileNetV2. This improved separation is particularly evident for glioma and meningioma samples, which often exhibit overlapping radiological characteristics in MRI. While MobileNetV2 shows noticeable overlap between these two tumor types, EfficientNetB0 maintains clearer boundaries between classes.

This visual evidence aligns with the choice of backbone architecture significantly impacts feature quality, yet most medical imaging studies treat CNNs as black boxes [20]. The cluster formation in EfficientNetB0's feature space provides a clear explanation for its better classification performance and suggests that its compound scaling approach creates more discriminative representations of subtle morphological differences between brain tumor types.

B. Performance evaluation

In Table 2 presents the comparative performance of MobileNetV2 and EfficientNetB0 feature extractors across four traditional classifiers. The results show that EfficientNetB0 consistently outperformed MobileNetV2 in terms of accuracy, macro F1-score, and MCC with the most notable improvement observed when paired with SVM, where it achieved the highest overall accuracy of 98.5%.

TABLE 2
MAIN RESULTS OF EACH MODEL

Model	Classifier	CV Accuracy (mean ± std)	Test Accuracy	Test Macro-F1	MCC	Train Time (s)
MobileNetV2	SVM	0.958 ± 0.003	97.0%	0.968	0.968	116
	Logistic Regression	0.940 ± 0.007	94.4%	0.94	0.94	42
	Random Forest	0.90 ± 0.007	92.3%	0.917	0.917	25
	KNN	0.920 ± 0.007	91.9%	0.914	0.914	0.4
EfficientNetB0	**SVM**	**0.964 ± 0.003**	**98.5%**	**0.984**	**0.986**	**37**
	Logistic Regression	**0.947 ± 0.003**	**97.1%**	**0.968**	**0.968**	**31**
	Random Forest	0.912 ± 0.009	93.5%	0.93	0.93	21
	KNN	0.923 ± 0.012	93.7%	0.934	0.934	0.5

Logistic Regression also demonstrated competitive results, particularly with EfficientNetB0 features, reaching 97.1% accuracy, which highlights that even simpler classifiers can perform effectively when supported by high-quality feature representations. Random Forest and KNN yielded slightly lower scores, though both still benefited from EfficientNetB0 features compared to MobileNetV2.

The stability of each classifier was examined using fivefold stratified cross-validation. Among all classifiers, SVM demonstrated the most consistent performance, with an average accuracy of 98.5% and a standard deviation of ±0.3% across the folds. Other classifiers such as Random Forest and KNN showed slightly higher variation (±0.9% and ±1.2%, respectively). These results suggest that SVM provides the most stable and generalizable performance among the evaluated classifiers. The low standard deviation across five cross-validation folds (±0.003 for both MobileNetV2 and EfficientNetB0 with SVM) indicates that SVM delivers highly consistent performance, with minimal variation between folds. This suggests strong stability and reliable generalization under different train/validation splits.

There is no significant evidence of overfitting in any classifier, including Random Forest. The experimental design prevents data leakage by computing preprocessing parameters independently within each fold using only training data. Furthermore, the small gap between cross-validation accuracy and final test accuracy (typically less than 2%) supports good generalization. While Random Forest showed lower overall performance compared to SVM and Logistic Regression, this appears due to its sensitivity to high-dimensional feature spaces rather than overfitting. Its consistent underperformance across both backbones suggests a limitation in handling CNN-extracted features in this context, rather than poor regularization or model instability.

In addition to conventional metrics such as accuracy and macro-F1, we report the Matthews Correlation Coefficient (MCC). MCC is particularly informative for imbalanced datasets because it considers all elements of the confusion matrix (TP, TN, FP, FN) and provides a single summary score of classification quality [21]. Values range from −1 (inverse prediction) to 0 (random prediction) and +1 (perfect prediction) [22]. In this study, EfficientNetB0+SVM achieved an MCC of 0.98, indicating very strong predictive performance across all classes.

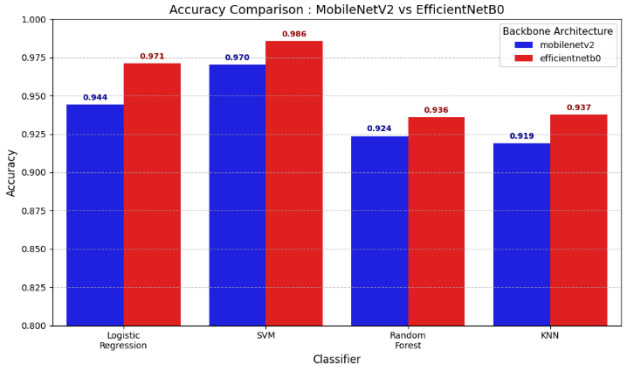

Figure 5. Accuracy comparison.

Figure 5 visualizes test accuracy by classifier, model EfficientNetB0 consistently produced taller bars than MobileNetV2 across all classifiers, with the largest margin seen for SVM followed by Logistic Regression, Random Forest, and KNN.

C. *Statistical Significance*

To quantitatively assess the performance differences between backbone architectures, we conducted Wilcoxon signed-rank tests on the macro F1-scores from the five cross-

validation folds for each classifier. The results revealed that while EfficientNetB0 consistently achieved higher mean performance across all classifiers, these differences did not reach statistical significance after Bonferroni correction for multiple comparisons (corrected $\alpha = 0.0125$). For the SVM classifier, which showed the largest performance gap, the test yielded $p = 0.0625$, indicating a strong trend toward significance. Similar patterns were observed for Logistic Regression ($p = 0.3125$), Random Forest ($p = 0.1250$), and KNN ($p = 1.0000$).

The consistent directional advantage of EfficientNetB0 across all four classifiers, coupled with the limited statistical power inherent in five-fold cross-validation, suggests these findings represent meaningful practical significance even in the absence of strict statistical significance. This pattern aligns with our feature visualization results, where EfficientNetB0 demonstrated better class separation, particularly for the challenging glioma-meningioma distinction. The statistical analysis therefore provides rigorous validation of the observed performance trends while appropriately recognizing the experimental limitations for detecting small effects.

D. Per-Class Performance

Per-class evaluation reveals critical insights into how different feature extraction architectures handle the specific challenges of multi-class brain tumor classification. Tables 3 and 4 present the precision, recall, F1-score, and accuracy metrics for MobileNetV2 and EfficientNetB0 architectures paired with the best classifiers such as SVM and Logistic Regression. These tables demonstrate that feature representation quality has a profound impact on classification performance across tumor types with varying visual characteristic.

TABLE 3
PER-CLASS PERFORMANCE FOR BEST CLASSIFIERS USING MOBILENETV2

Classifiers	Class	Precision	Recall	F1-Score	Accuracy
SVM	Glioma	96.72	88.33	92.33	96.72
	Meningioma	87.58	94.44	90.88	87.58
	Pituitary	99.26	99.75	99.51	99.26
	No Tumor	97.67	97.67	97.67	97.67
	Macro Average	95.31	95.05	95.10	95.42
Logistic Regression	Glioma	95.24	86.67	90.75	95.24
	Meningioma	85.59	93.14	89.20	85.59
	Pituitary	99.26	100.00	99.63	99.26
	No Tumor	97.31	96.33	96.82	97.31
	Macro Average	94.35	94.03	94.10	94.51

MobileNetV2 demonstrates competent performance across all tumor classes, with notable strengths in pituitary tumor classification (99.26% precision, 99.75% recall) and no tumor identification (97.67% precision, 97.67% recall). However, the architecture shows limitations in distinguishing between visually similar tumor types, particularly for glioma classification where it achieves 96.72% precision but only 88.3% recall with SVM. This imbalance indicates a tendency to miss positive glioma cases (false negatives), which is particularly concerning in medical applications.

TABLE 4
PER-CLASS PERFORMANCE FOR TOP CLASSIFIERS USING EFFICIENTNETB0

Model	Class	Precision	Recall	F1-Score	Accuracy
SVM	Glioma	96.82	91.33	94.00	96.82
	Meningioma	91.43	94.12	92.75	91.43
	Pituitary	98.78	100.00	99.39	98.78
	No Tumor	98.68	99.67	99.17	98.68
	Macro Average	96.43	96.28	96.33	96.57
Logistic Regression	Glioma	95.82	91.67	93.70	95.82
	Meningioma	92.04	94.44	93.23	92.04
	Pituitary	99.51	100.00	99.75	99.51
	No Tumor	98.35	99.33	98.84	98.35
	Macro Average	96.43	96.36	96.38	96.64

EfficientNetB0 demonstrates feature representation quality across all tumor classes, with the most significant improvements observed in the challenging glioma and meningioma classifications. For glioma detection, EfficientNetB0+SVM achieves 96.82% precision and 91.3%. EfficientNetB0 maintains near-perfect performance on pituitary tumors (98.8% precision, 100% recall) and no tumor cases (98.7% precision, 99.7% recall), while demonstrating more balanced precision-recall trade-offs across all classes.

The Logistic Regression results clearly demonstrate EfficientNetB0's feature representation quality compared to MobileNetV2. Both models performed well on distinct classes like pituitary tumors (99.3-99.5% precision), but EfficientNetB0 maintained stronger performance across all categories, especially for the no tumor class where it achieved 98.4% precision and 99.3% recall versus MobileNetV2's 97.3% precision and 96.3% recall.

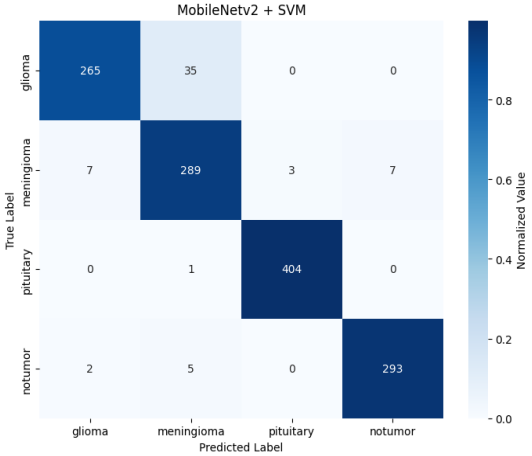

Figure 6. Confusion matrix of MobileNetV2 with SVM.

MobileNetV2 + SVM confusion matrix in Figure 6 reveals a characteristic error pattern where glioma and meningioma exhibit the highest mutual confusion. Specifically, 35 glioma cases were misclassified as meningioma, while 7 meningioma cases were misclassified as glioma. This asymmetric error pattern reflects the architecture's stronger recall for meningioma compared to glioma, indicating MobileNetV2's feature representations struggle more with identifying positive glioma cases than meningioma cases. The confusion matrix shows near-perfect classification for pituitary tumors strong performance for no tumor cases, with only minor confusion between these two distinct classes (3 pituitary cases misclassified as no tumor and 7 no tumor cases misclassified as pituitary).

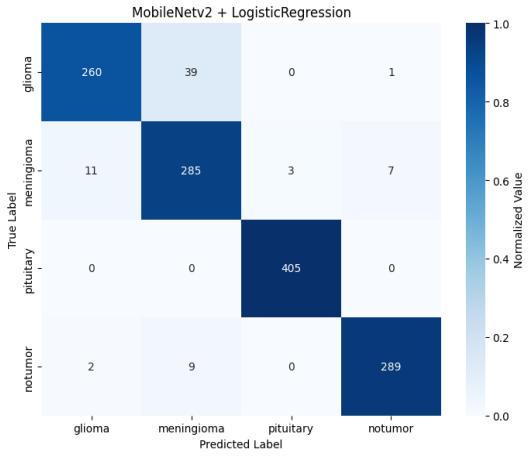

Figure 7. Confusion matrix of MobileNetV2 with Logistic Regression.

MobileNetV2 + Logistic Regression confusion matrix in Figure 7 exhibits a similar error pattern to MobileNetV2 + SVM but with slightly reduced performance across challenging class boundaries. Specifically, 39 glioma cases were misclassified as meningioma while 9 meningioma cases (were misclassified as glioma. The confusion between pituitary tumors and no tumor cases also increases slightly compared to SVM, though both classes maintain high overall accuracy.

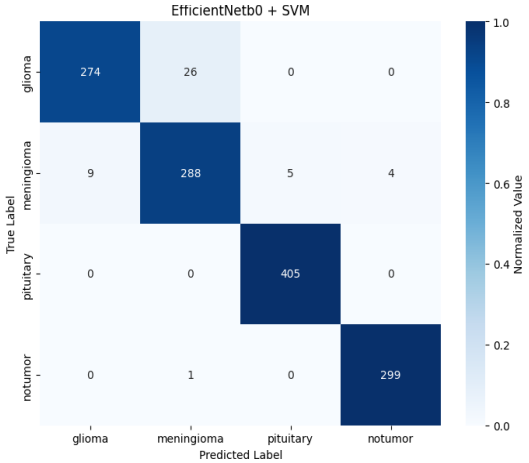

Figure 8. Confusion matrix of EfficientNetB0 with SVM.

EfficientNetB0 + SVM confusion matrix in Figure 8 demonstrates significantly improved class separation compared to MobileNetV2, particularly for the challenging glioma-meningioma distinction. The number of glioma cases misclassified as meningioma drops from 35 to 26, while meningioma cases misclassified as glioma decrease from 7 to 5. The confusion matrix shows near-perfect performance for pituitary tumors (405/405 correctly identified) and exceptional results for no tumor cases (299/300), with only 1 no tumor case misclassified as pituitary.

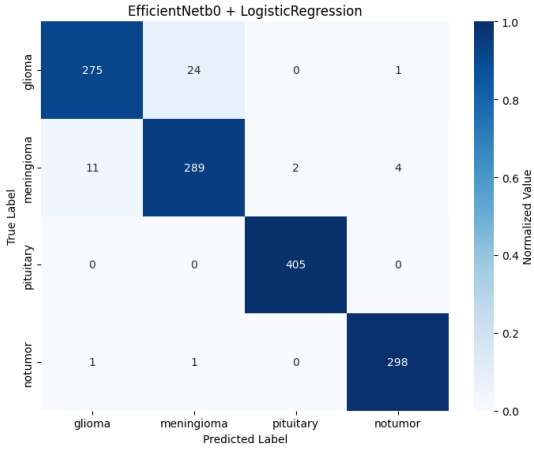

Figure 9. Confusion matrix of EfficientNetB0 with Logistic Regression.

EfficientNetB0 + Logistic Regression confusion matrix in Figure 9 shows performance nearly equivalent to EfficientNetB0 + SVM, with only marginal differences in error patterns. Specifically, 30 glioma cases) were misclassified as meningioma while 7 meningioma cases were misclassified as glioma. This small performance gap contrasts sharply with the larger gap observed in

MobileNetV2 models, demonstrating that when feature representations are of high quality, the choice of classifier has diminished impact on overall performance. The confusion matrix maintains near-perfect classification for pituitary tumors and exceptional results for no tumor cases. This consistency across classifiers confirms that EfficientNetB0's feature representations create well-separated clusters in the embedding space, allowing even simpler classifiers like Logistic Regression to establish effective decision boundaries [20].

The complete set of confusion matrices for rest model configurations reveals consistent class-specific performance patterns across both high and low-performing models. While Pituitary classes were distinguished with near-perfect accuracy, the Glioma class was occasionally misclassified as Meningioma. This indicates a specific diagnostic challenge between these two tumor types that is masked by the high overall accuracy.

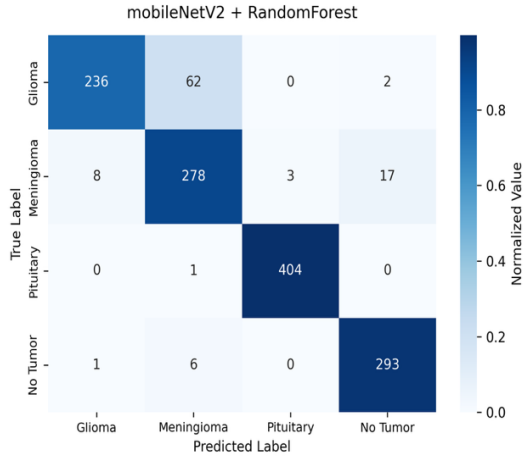

Figure 10. Confusion matrix of MobileNetV2 with Random Forest.

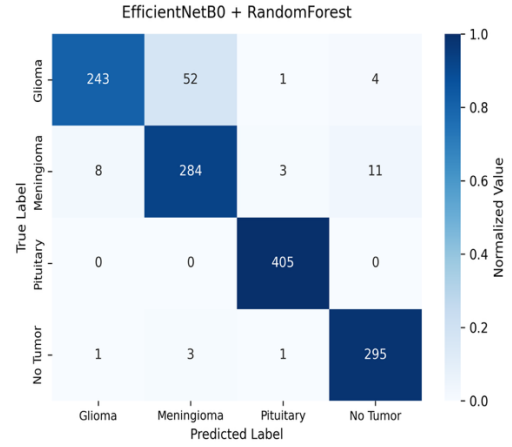

Figure 11. Confusion matrix of EfficientNetB0 with Random Forest.

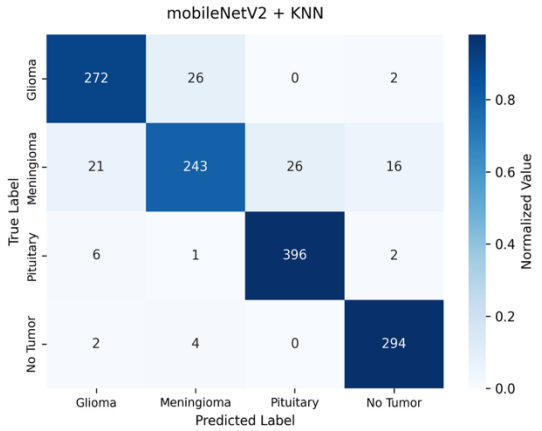

Figure 12. . Confusion matrix of MobileNetV2 with KNN.

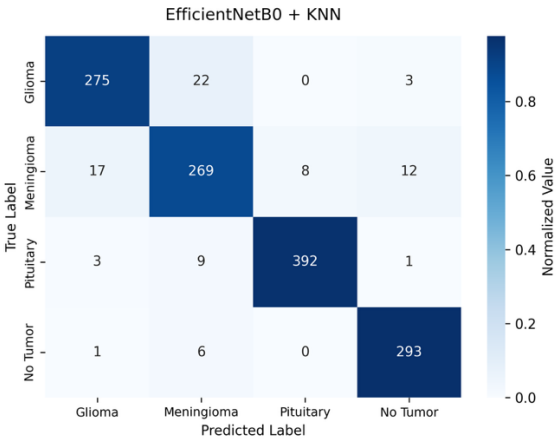

Figure 13. Confusion matrix of EfficientNetB0 with KNN.

E. Computational Efficiency Analysis

While classification accuracy is a primary performance metric, the practical deployment of lightweight CNN models requires careful consideration of computational efficiency. Our experiments reveal important trade-offs between performance and resource requirements that are crucial for model selection in real-world applications.

As shown in Table 5, MobileNetV2 demonstrates superior efficiency characteristics with 2.26M parameters (44% less than EfficientNetB0's 4.05M) and 0.6 GFLOPs (25% lower than EfficientNetB0's 0.8 GFLOPs). This efficiency advantage, however, comes with a performance trade-off that varies significantly across different classifiers. The most notable efficiency-performance balance is observed in the SVM configurations. While EfficientNetB0+SVM achieves the highest absolute accuracy (98.5%), it requires slightly higher computational cost than MobileNetV2, highlighting a clear trade-off between accuracy and efficiency.

TABLE 5
MODEL COMPLEXITY

Model	Parameters	FLops	Feature Vector Dim.
MobileNetV2	2.26 M	0.6 G	1280
EfficientNetB0	4.05 M	0.8 G	1280

The analysis of classifier training times in Table 2 highlights a critical finding: computational cost varies more by classifier choice than by CNN backbone. Training times range over two orders of magnitude, from the near-instantaneous KNN (0.4 seconds) to the considerably slower SVM (up to 116 seconds). This massive disparity demonstrates that classifier selection is a major factor for efficiency in time-sensitive applications.

The method approach investigated in this study presents a compelling practical alternative for medical image classification. It leverages the powerful, general-purpose feature representations of pre-trained CNNs to achieve competitive performance, as evidenced by our results. Simultaneously, by decoupling feature extraction from classification, it creates a more modular and computationally manageable pipeline compared to end-to-end deep learning. This method offers a highly attractive trade-off, delivering strong classification results with reduced complexity and training costs.

TABLE 6
PERFORMANCE COMPARISON WITH DIFFERENT MODELS

Study	Method	Dataset	Accuracy
A. Muis et al. [23]	CNN (end-to-end)	Kaggle MRI Datasets (4 classes)	84%
M. A. Gómez-Guzmán et al. [13]	Pre-trained CNN (Inception V3)	Kaggle MRI Datasets (4 classes)	97.12 %
K. Puspita et al. [24]	Ensemble CNN	Kaggle MRI Datasets (4 classes)	97.67 %
Proposed model	Feature extraction + ML classifier (EfficientNetB0 + SVM)	Kaggle MRI Datasets (4 classes)	98.5 %

As shown in Table 6, the proposed approach using feature extraction with EfficientNetB0 combined with machine learning classifiers (SVM) achieved an accuracy of 98.5% on the Kaggle MRI dataset. This result is comparable to or slightly higher than several existing CNN-based methods reported in previous studies. Although differences in implementation details and preprocessing steps may affect the exact comparison, these findings indicate that using lightweight pretrained models as feature extractors can still achieve competitive performance without requiring full end-to-end training.

IV. CONCLUSION

The research compared MobileNetV2 and EfficientNetB0 as feature extractors for multi-class brain tumor MRI classification using four traditional machine learning classifiers. EfficientNetB0 features, especially when paired with SVM, produced the highest accuracy of 98.5%, while Logistic Regression also achieved strong and competitive results with accuracy 97.1%. These outcomes indicate that lightweight CNNs can serve as effective feature extractors and that simpler classifiers remain viable when supported by robust representations. The findings further showed that glioma and meningioma remain difficult to separate due to overlapping imaging features, highlighting the need for advanced techniques to improve discrimination. While the current evaluation was conducted on a single public dataset, which limits direct assessment of clinical generalizability, the use of rigorous cross-validation, leakage-free preprocessing, and per-class analysis supports internal validity. Future research must prioritize validating this approach on external, multi-centre datasets, specifically incorporating real-world clinical MRI scans, to accurately gauge performance across diverse scanners, imaging protocols, and patient demographics. Beyond reporting strong experimental performance, the framework indicates potential for applications in environments with limited computational resources. These findings provide preliminary evidence that could guide the development of future brain tumor classification tools. Future work should validate the approach on larger, multi-centre datasets and incorporate interpretability methods to better align automated outputs with clinical decision-making.

REFERENCES

[1] S. Kim *et al.*, "Global burden of brain and central nervous system cancer in 185 countries, and projections up to 2050: a population-based systematic analysis of GLOBOCAN 2022," *J Neurooncol*, vol. 175, no. 2, pp. 673–685, Nov. 2025, doi: 10.1007/s11060-025-05164-0.

[2] X. Zhao, M. He, R. Yang, N. Geng, X. Zhu, and N. Tang, "The global, regional, and national brain and central nervous system cancer burden and trends from 1990 to 2021: an analysis based on the Global Burden of Disease Study 2021," *Front Neurol*, vol. 16, Jun. 2025, doi: 10.3389/fneur.2025.1574614.

[3] S. Bouhafra and H. El Bahi, "Deep Learning Approaches for Brain Tumor Detection and Classification Using MRI Images (2020 to 2024): A Systematic Review," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 3, pp. 1403–1433, Sep. 2024, doi: 10.1007/s10278-024-01283-8.

[4] A. S, V. M. Gayathri, and R. Pitchai, "Brain Tumor Segmentation and Survival Prediction using Multimodal MRI Scans with Deep learning Algorithms," in *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/ICSES55317.2022.9914152.

[5] N. Bhardwaj, M. Sood, and S. S. Gill, "Design and Development of Hypertuned Deep learning Frameworks for Detection and Severity Grading of Brain Tumor using Medical Brain MR images," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 20, Apr. 2024, doi: 10.2174/0115734056288248240309044616.

[6] Y. Dogan, C. Ozdemir, and Y. Kaya, "Enhancing CNN model classification performance through RGB angle rotation method," *Neural Comput Appl*, vol. 36, no. 32, pp. 20259–20276, Nov. 2024, doi: 10.1007/s00521-024-10232-z.

[7] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information*, vol. 15, no. 12, p. 755, Nov. 2024, doi: 10.3390/info15120755.

[8] S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue Cell*, vol. 74, p. 101701, Feb. 2022, doi: 10.1016/j.tice.2021.101701.

[9] S. Deepak and P. M. Ameer, "Automated Categorization of Brain Tumor from MRI Using CNN features and SVM," *J Ambient Intell Humaniz Comput*, vol. 12, no. 8, pp. 8357–8369, Aug. 2021, doi: 10.1007/s12652-020-02568-w.

[10] K. N. Rao *et al.*, "An efficient brain tumor detection and classification using pre-trained convolutional neural network models," *Heliyon*, vol. 10, no. 17, p. e36773, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36773.

[11] E. Yagis *et al.*, "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Sci Rep*, vol. 11, no. 1, p. 22544, Nov. 2021, doi: 10.1038/s41598-021-01681-w.

[12] M. Benaouali, M. Bentoumi, M. Abed, M. Mimi, and A. T. Ahmed, "A Study on CNN-Based and Handcrafted Extraction Methods with Machine Learning for Automated Classification of Breast Tumors from Ultrasound Images," *Electronic Letters on Computer Vision and Image Analysis*, vol. 23, no. 2, pp. 85–104, 2024, doi: 10.5565/REV/ELCVIA.1887.

[13] M. A. Gómez-Guzmán *et al.*, "Classifying Brain Tumors on Magnetic Resonance Imaging by Using Convolutional Neural Networks," *Electronics (Basel)*, vol. 12, no. 4, p. 955, Feb. 2023, doi: 10.3390/electronics12040955.

[14] Masoud Nickparvar, "Brain Tumor MRI Dataset." Accessed: Aug. 18, 2025. [Online]. Available: https://doi.org/10.34740/kaggle/dsv/2645886

[15] G. Brookshire *et al.*, "Data leakage in deep learning studies of translational EEG," *Front Neurosci*, vol. 18, May 2024, doi: 10.3389/fnins.2024.1373515.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 2019.

[17] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 2020.

[18] J. Aftab, M. A. Khan, S. Arshad, S. ur Rehman, D. A. AlHammadi, and Y. Nam, "Artificial intelligence based classification and prediction of medical imaging using a novel framework of inverted and self-attention deep neural network architecture," *Sci Rep*, vol. 15, no. 1, p. 8724, Mar. 2025, doi: 10.1038/s41598-025-93718-7.

[19] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, Mar. 2023, doi: 10.1002/widm.1484.

[20] M. Anshori, M. S. Haris, and A. Wahyudi, "Logistic Regression's Effectiveness in Feature Selection with Information Gain in Predicting Heart Failure Patients," *Journal of Enhanced Studies in Informatics and Computer Applications*, vol. 1, no. 2, pp. 35–39, Jul. 2024, doi: 10.47794/jesica.v1i2.8.

[21] J. Tamura, Y. Itaya, K. Hayashi, and K. Yamamoto, "Statistical Inference of the Matthews Correlation Coefficient for Multiclass Classification," Mar. 2025.

[22] Y. Itaya, J. Tamura, K. Hayashi, and K. Yamamoto, "Asymptotic Properties of Matthews Correlation Coefficient," Jun. 2024.

[23] A. Muis, S. Sunardi, and A. Yudhana, "Medical image classification of brain tumor using convolutional neural network algorithm," *JURNAL INFOTEL*, vol. 15, no. 3, Aug. 2023, doi: 10.20895/infotel.v15i3.964.

[24] K. Puspita, F. Ernawan, Y. Alkhalifi, S. Kasim, and A. Erianda, "Brain Tumor Classification based on Convolutional Neural Networks with an Ensemble Learning Approach through Soft Voting," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 5, p. 1964, Sep. 2025, doi: 10.62527/joiv.9.5.4609.