Scalable and Efficient Multi-Class Brain Tumor Classification with a Compact Hybrid Deep Learning Model for Real-Time Applications

Sohaib R. Awad[1], Amar I. Daood[2†] and Akram A. Dawood[2]

[1]Department of Computer and Information Engineering, Ninevah University, Mosul, Iraq

[2]Department of Computer Engineering, College of Engineering, University of Mosul, Mosul, Iraq

***Abstract*—Medical diagnostics require brain tumor classification to operate in real-time so the task demands accurate results with efficient processing abilities. A new hybrid deep learning solution merges convolutional neural networks (CNNs) with support vector machines (SVMs) to improve classification results as this paper describes. A total of four tumor categories including glioma, meningioma, and pituitary tumors together with no tumor appearance contribute to the magnetic resonance imaging (MRI) dataset are used for analysis. We applied and organized three pre-trained deep learning models: AlexNet, DarkNet-19, and ResNet-50 for comparison. A newly engineered compact CNN model linked with an SVM classifier brought decreased model dimensions while keeping excellent accuracy rates. A proposed compact CNN model delivers 97.50% accuracy through its smaller 2.38 MB size and an additional SVM integration results in 97.45% accuracy using 1.43 MB. A Graphical User Interface (GUI) system comprising automated tumor classification capabilities is created to improve real-time systems that visualize MRI scans and illustrate predicted labels in addition to displaying confidence scores. A GUI enables smooth access to the trained model while being suitable for medical practice mobile healthcare environments and edge computing needs. The proposed system shows that lightweight architectures work excellently in real-time system applications especially when used for edge computing and mobile healthcare frameworks. The proposed solution demonstrates superiority over established models through its ability to scale efficiently.**

Index Terms*—Brain tumor classification, Deep learning, Medical diagnostics, Real-time applications, Scalable AI solutions.

I. Introduction

Brain tumor classification requires an early and precise identification for medical diagnostic purposes. The diagnostic process together with treatment recommendations becomes more efficient for radiologists when they use magnetic resonance imaging (MRI)-based classification methods. Deep learning technology especially Convolutional Neural Networks (CNNs) delivered significant improvements to brain tumor classification precision during recent times. The high computational requirements of advanced CNN models render them unusable when running real-time or edge-computing operations. Early identification and treatment of such tumors are significant for patient outcomes, and the classification of brain tumors in medical diagnostics is a critical task (Sharma, et al., 2023). It is very difficult to separate true tumor structures, namely, glioma, meningioma, pituitary, and no tumor due to the complexity of the tumor structures and variation of medical image data (Ullah, et al., 2023). As the demand for radiologists using automated systems to assist them continues to grow, there is greater demand for solutions that are not only very accurate, but also computationally efficient, and deployable in resource-constrained environments (Mohammed, et al., 2024; Jana, et al., 2023). Due to the ability to learn hierarchical and complex features directly from raw medical imaging data, deep learning has become an optimal solution for brain tumor classification (Awad, et al., 2022; Ganaie, et al., 2022). Among these, CNNs in particular have demonstrated exceptionally high accuracy in achieving a wide range of different tasks such as pattern classification (Daood, Al-Saegh and Mahmood, 2023; Abdulaziz and Dawood, 2023), smart surveillance (Mohammed and Daood, 2021), biometric (Alhafidh, Hagem and Daood, 2022), smart home (Alhafidh, Daood and Allen, 2018), and medical image analysis (Müller, Soto-Rey and Kramer, 2022; Jasim, et al., 2021). Nevertheless, conventional deep learning models are prone to their size being huge and having high computational requirements (Akinbo and Daramola, 2021; Abdullah, Mohammed and Awad, 2024) rendering them impractical for real-time and edge computing applications. Research in brain tumor classification through deep learning techniques makes progress daily according to two recent studies because it enables automated tumor detection and enhances segmentation quality (Al-Jammas,

ARO-The Scientific Journal of Koya University
Vol. XIII, No.1(2025), Article ID: ARO.12017. 13 pages
DOI: 10.14500/aro. 12017
Received: 22 January 2025; Accepted: 19 April 2025
Regular research paper; Published: 03 May 2025
†Corresponding author's e-mail: amar.daood@uomosul.edu.iq
Copyright © 2025 Sohaib R. Awad, Amar I. Daood, and Akram A. Dawood. This is an open-access article distributed under the Creative Commons Attribution License (CC BY-NC-SA 4.0).

et al., 2024). Feature extraction capabilities of CNN-based models received development through new model implementations and multiple CNN strategies worked effectively for ensemble-based MRI tumor detection tasks. The advancement demonstrates why the lightweight design of neural network architecture requires both accurate results and efficient computing power and instant detection capability to achieve effective tumor recognition (Al-Mukhtar, et al., 2024; Santoso, Supriyono and Utama, 2024).

We present scalable and efficient models addressing the above restrictions in this study. This research implements a minimal CNN model that operates with an support vector machine (SVM) classification system to cope with this challenge. The hybrid CNN-SVM method optimizes performance efficiency and decreases model complexity despite using deep feature learning methods independently of traditional deep learning methods. The research study tests three deep learning frameworks, namely, ResNet-50, Alex-Net, and DarkNet-19 alongside the constructed compact CNN-SVM model. Our contributions include:

1. A compact CNN model development for medical image classification presents efficient computations while reducing expenses.
2. The combination of CNN with SVM classifiers improves system efficiency without affecting the accuracy of results.
3. The study analyses deep learning models based on their measurement performance alongside their speed capabilities and time responsiveness.
4. The study demonstrates the relationship between important performance aspects between state-of-the-art methods by showing how model precision interacts with operational efficiency.

The paper advances present research by prioritizing the equilibrium between precise results and low computational demands needed for quick medical use scenarios. The rest of this paper is organized as follows: In section 2, we present a review of related works, by summarizing the latest developments in brain tumor classification. In section 3, the methodology is presented with an overview of dataset preparation, model design, and training strategies. In section 4, we examine experimental results, and in section 5, we compare our work with other recent state-of-the-art approaches. Finally, section 6 concludes with the study and possible directions for future work.

II. Literature Review

Classifying brain tumors is an important task in medical diagnostics that typically demands powerful methods to achieve high accuracy, while also satisfying constraints when being deployed in environments whose resources are limited (Ramanagiri, Mukunthan and Balamurugan, 2024; Rasool, et al., 2024). Recent advancements in artificial intelligence, particularly in deep learning and hybrid approaches, have significantly contributed to the progress in this domain (Onuiri, John and Umeaka, 2024; Mavaddati, 2024). This section reviews these key studies that have informed and influenced the present research.

Transfer learning has been a very powerful tool in medical image analysis as it utilizes pre-trained models for application on problems where annotated data are scarce. ResNet-50 was used in one study (Sahaai, et al., 2022) where they used it for brain tumor classification and illustrated its efficacy in producing intricate features from MRI image. (Srinivas, et al., 2022) investigated deep transfer learning approaches for brain tumor classification through analyzing the robustness of models ResNet-50, Inception-v3, and VGG-16 in medical imaging tasks. However, these studies mainly pushed for high accuracy, often ignoring constraints regarding the size of the model and computational efficiency.

Recent research has been performed to look at what it means to use deep learning feature extractors with traditional machine learning (ML) classifiers to improve computational efficiency and accuracy. Another study (Karim, Mahmood and Sah, 2023) fine-tuned a deep transfer learning model with an SVM that produced better performance than previous confidential. In a similar vein, another approach in (Yadav, et al., 2024) suggested a modified ResNet-50 model and worked well, yet at the cost of expensive computation as a result of model size. Nevertheless, the actual deployment of these hybrid frameworks in resource-constrained environments is still limited.

Large pre-trained models are limited in their scalability; hence, compact architectures have become a hot topic for their existence and scalability of inferencing and also to their ability to equip the real-time application. In one research effort, a deep transfer learning framework for multi-class brain tumor classification (Divya, Suresh and John, 2020) obtained promising results and showed the necessity for lightweight models. A related work (Shamshad, et al., 2024) considered improvements in the efficacy of transfer learning methods, for example, models such as VGG-19, VGG-16, ResNet-50, Mobile-Net, Inception-v3, and Dense-Net, on MRI datasets, noting the difficulties in achieving high accuracy at the same time as low model size and computational cost. A spatial pyramid pooling layer was added to another study (Neamah, et al., 2024) which introduced a modified ResNet-50 model trained with pre-trained parameters to combat overfitting, and partially addresses scalability concerns with high accuracy. The present studies are useful in understanding these trade-offs between model complexity and deployment feasibility, and correspond with the intended aims of the present research.

Multiple research groups have conducted studies of deep learning for brain tumor classification while focusing on enhancing accuracy levels. Department of Medical Sciences Dalhousie University and author Marie-Claude Arnaud explore the effectiveness of transfer learning with ResNet-50, Inception-v3, and VGG-16 for brain tumor identification while acknowledging their high computational costs (Sharma, et al., 2023; Ullah, et al., 2023). Hand-in-hand integration between CNNs and SVM ML classifiers leads to model size reduction without any loss in classification effectiveness (Karim, Mahmood and Sah, 2023). The evaluation of compact architectures suitable for real-time deployment is not addressed in several studies that currently exist. Research

http://dx.doi.org/10.14500/aro.12017

nowadays strives to develop lightweight architectures for brain tumor classification with reduced computational expenses and stable robustness. The research brings forward an optimized compact CNN model for brain tumor classification alongside extensive performance evaluation.

Many previous studies have made impressive progress in terms of accuracy; however, these works often fail to account for important elements such as model scalability, computational efficiency, and practical feasibility of real-time deployment. ResNet-50, VGG-16, and VGG-19 are extremely correct but they take up a great deal of resources and are consequently insufficient for use in mobile or edge devices. Such hybrids promise to address these issues, but usually, this comes at the cost of more complex frameworks, which increases computational overhead. On the other hand, compact architectures open the door to scalability, but more optimization is needed to close the gap on larger model accuracy.

To address these gaps, this study introduces a new compact CNN model, and hybrid frameworks that combine using CNNs with SVM classifiers. In the next parts, we will show that the proposed models result in competitive accuracy compared with baseline models, with a significant reduction in model size and classification time, which makes them practical for deployment in resource-constrained environments. This research builds on and exceeds the limitations recognized in the literature in brain tumor classification, presenting scalable and efficient AI-powered solutions.

III. Materials and Methods

This paper is an attempt to develop an efficient and accurate classification system for brain tumor prediction, which incorporates deep learning and ML approaches. The multiple stages involved in the methodology include dataset preparation, model training, feature extraction, and hybrid classifier integration. The design of each stage was intended to guarantee high accuracy, a low computational cost, and practical applicability in real-world scenarios.

A. Overview of the Proposed Method

This research focuses on classifying brain tumor MRI images into four classes: Included are glioma, meningioma, no tumor, and pituitary. For this, three pre-trained networks are chosen: Alex-Net (depth 8), DarkNet-19 (depth 19), and ResNet-50 (depth 50). We chose these models because they have an ascending architectural complexity and have exhibited good performance on a diversity of datasets. However, due to the large model size, VGG-16 and VGG-19, though with similar accuracy, were omitted, and they are not fit for real-time applications and deployment in low-resource environments.

We evaluated two optimizers – Adaptive Moment Estimation (ADAM) and Stochastic Gradient Descent with Momentum (SGDM) – on every network to tune out optimal hyper-parameters in pursuit of the best accuracy versus performance trade-off. To improve deployment suitability for the further scope, the original fully connected dense layers were removed, followed by passing the extracted features to ML models, in particular SVM, to ascertain the extent of identifying the best accuracy and performance trade-off within the image classification. Prolonged experimentation with various classifiers demonstrated that the SVM approach is mutually robust for high-dimensional feature spaces. In the end, a compact ML model was proposed aiming to strike a balance between classification accuracy, classification latency, and model size, thus offering a strong candidate for real-time use in an edge deployment scenario.

B. Dataset Preparation

A comprehensive dataset of 10,183 MRI images was compiled from three reputable sources: Figshare, SARTAJ, and Br35H. These images are classified into four categories: Other benign tumors such as meningioma, glioma, no tumor, and pituitary tumors (Mohammed, et al., 2024). To ensure uniformity and compliance with deep learning architectures, each image was normalized to a resolution of 256×256 pixels. The dataset was divided evenly between training and validation (80%), and testing (20%) functions to achieve balanced learning together with generalization. Model parameters received optimization through the training data and the validation data subset from training performed hyper-parameters alteration and performance checking to prevent overfitting. The model works properly on new data because it uses the validation set performance evaluation to ensure good generalization before final testing. The model testing process utilized a distinct 20% segment of data as a testing set which remained separate throughout for objective evaluation of classification capabilities. Table I demonstrates an equal proportion of classes throughout the data distribution.

The model achieved generalization by utilizing data augmentation through rotational techniques at ±45° angles together with flipping operations. Augmentations for these remained in semantic integrity within the image, yet they varied to help provide the model's ability to generalize to unseen data. To conform to pre-trained models, input pre-processing resized images for compatibility with Alex-Net (227×227×3), DarkNet-19 (256×256×3), and ResNet-50 (224×224×3). To satisfy these requirements we converted grayscale images into RGB format by duplicating the single channel across 3 dimensions. Fig. 1 shows class samples of the dataset.

TABLE I
Dataset Class Distribution

Class	Number of images	Training/validation (80%)	Testing (20%)
Glioma	2,547	2,038	509
Meningioma	2,582	2,065	517
No tumor	2,396	1,917	479
Pituitary	2,658	2,126	532
Total	10,183	8,146	2,037

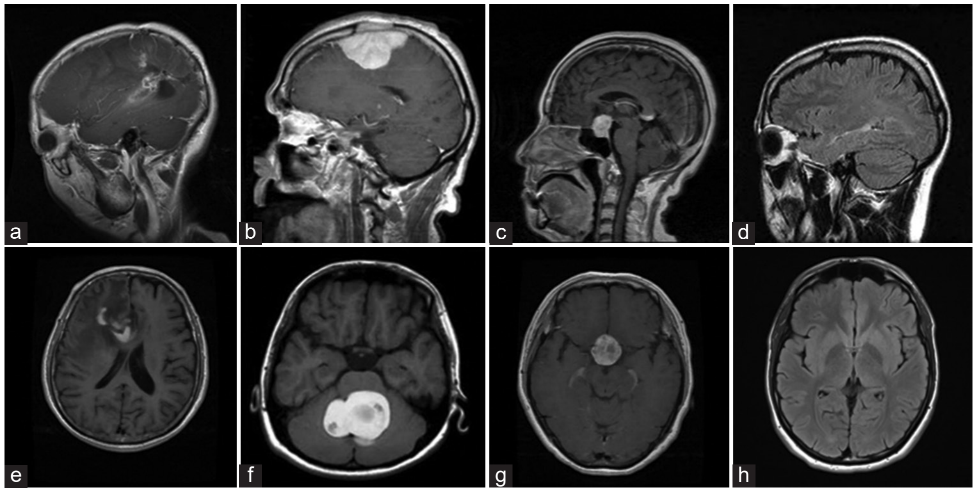

Fig. 1. Sample images from each dataset class. (a) Glioma, (b) Meningioma, (c) Pituitary, (d) No tumor, (e) Glioma, (f) Meningioma, (g) Pituitary, (h) No tumor.

C. Pre-trained Networks and Optimization

Three pre-trained models were selected: To understand the recent success of deep neural networks, we study the factors that contribute to learning depth in three state-of-the-art network models: Alex-Net, DarkNet-19, and ResNet-50, which have depths of 8, 19, and 50, respectively. We use these models to investigate the trade-offs between accuracy, complexity, and computational efficiency. Exclusion of VGG-16 and VGG-19 was also due to unnecessary large size, making them impractical for real-time deployment. The optimal hyper-parameters and best accuracy for each architecture were identified using each model trained by two different optimization strategies, ADAM and SGDM. The pre-trained network specifications are summarized in Table II.

This training entailed tuning hyper-parameters, such as learning rate, epoch count, and batch size. We also train models with ADAM and use an initial learning rate of 0.001 and a learning rate drop factor of 0.5 every 10 epochs. We used a similar learning rate schedule as SGDM, but including momentum to stabilize training. The balance and effectiveness of each model were evaluated by computing performance metrics: Accuracy, Precision, Recall, and F1-score. All networks standardize MATLAB training settings (Table III).

TABLE II
Summary of Pre-trained Models

Model	Total layers	Connections	Depth	Learnable parameters	Input image size
Alex-Net	25	24	8	59.8 M	227×227×3
DarkNet-19	65	64	19	20.3 M	256×256×3
ResNet-50	177	192	50	25.6 M	224×224×3

TABLE III
Training Configuration and Parameters

Parameter	Value
Initial learning rate	0.001
Learning rate drop factor	0.5
Learning rate drop period	10 epochs
Maximum epochs	30
Batch size	32
Validation frequency	50
Hardware	GPU: NVIDIA RTX 3060 with 6 GB VRAM. Processor: Intel Core i7 12th Gen. Memory: 16 GB RAM

D. Hybrid Frameworks with Feature Extraction

To decrease the size of the model by eliminating the dense (fully connected) layers, and simplify the classification time, while maintaining high accuracy, models were pre-trained. Features were extracted from the last convolutional layers of the trained models: The number of features in the 3 models is Alex-Net (4096 features), ResNet-50 (2048 features), and DarkNet-19 (1000 features). Using these features, ML classifiers were trained with both logistic regression and standard SVM with an SVM box constraint of 5, and the kernel is Radial Basis Function RBF performing best after hyper-parameter tuning. Random forest, decision tree, and XG-Boost classifiers were also compared for comparison

This hybrid framework uses CNN features to retain the discriminative power of CNN features, but lowering the overall model complexity. Feature dimensionality was reduced without losing significant information by using Global Average Pooling (GAP) before feature extraction. Classification times were considerably improved and the disk storage requirement was reduced with the hybrid approach.

However, removing dense layers offered challenges on the account of reduced learnable parameters, as well as the need for robust feature representations to achieve the same level of accuracy as the baseline model (Kokhazadeh, et al., 2024; Awad and Alghareb, 2025). Careful management was required for the dimensionality of extracted features to prevent computational overhead (Xia, et al., 2018; Pan, et al., 2020). To obtain the best performance, we then fine tune the hyper-parameters, yet more crucially, we also needed to ensure the compatibility between CNN feature extractors, and SVM classifiers. Mitigation of these challenges is shown using GAP, and rigorous evaluation of hybrid models.

E. Proposed Compact CNN Model

Our proposed model consists of a compact CNN architecture with:

- Input layer: 128×128 RGB image resolution
- Four convolutional layers with batch normalization
- GAP layer for feature extraction
- SVM classifier for final classification.

Using CNN-derived vectors the feature extraction process feeds information to an SVM classifier that improves performance and decreases operational complexity. To balance accuracy and computational efficiency, a compact CNN model was proposed. This model was lighter in terms of layers and learnable parameters, being smaller. Instead of using dense layers, the architecture used GAP to obtain 256 features that were fed into an SVM classifier. With significantly faster classification times and smaller model size, the compact model was competitive to its full counterpart, and hence suitable for real-time edge computing and mobile applications. The proposed ML model architecture specifications are presented in Table IV.

To further enhance the deployment efficiency of the compact CNN, the dense layer was removed, and the model was integrated with an SVM classifier. The features extracted from the last convolutional layer (256 features) were used as input to the SVM, as shown in Fig. 2 below.

F. Models Evaluation

Accuracy, precision, recall, and F1-score were evaluated (equations 1–4) along with classification time, size of the model, and number of learnable parameters of each model (Naidu, Zuva and Sibanda, 2023; Obi, 2023). Results showed a trade-off between efficiency and accuracy, where the proposed compact CNN model provides an optimal balance. The results section presents detailed discussions of these metrics and their implications, as well as tables summarizing comparative findings.

TABLE IV
Architecture Overview of the Proposed Compact CNN Model

Attribute		Value
Input image size, Optimizer		128×128×3, ADAM
Total layers, Number of connections, Depth		19, 18, 5
Number of convolutional layers	4	conv_1 128×128×32, conv_2 64×64×64, conv_3 32×32×128, conv_4 16×16×256
Number of fully connected layers		1
Pooling layers		3
Number of batch normalizations		4
Number of activation layers		4
Number of classes	4	Glioma, Meningioma, Pituitary, and No tumor
Total learnable parameters		651.5 K
Fully connected layer parameters		262.15 K

$$Accuracy\ \% = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (1)$$

$$Precision\ \% = \frac{TP}{TP+FP} \times 100 \quad (2)$$

$$Recall\ \% = \frac{TP}{TP+FN} \times 100 \quad (3)$$

$$F1-Score\ \% = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100 \quad (4)$$

The evaluation metrics are based upon true positive (TP), which is the number of correctly predicted positive instances; true negative (TN), which is the number of correctly predicted negative instances; false positive (FP), which is the number of instances incorrectly classified as positive; and false negative (FN), which is the number of instances incorrectly classified as negative (Raja, et al., 2024). These parameters serve as bases for the calculation of critical performance metrics such as reliability and cycle time, to serve as an overall metric of the predictive capabilities of the model.

IV. Experimental Results

In this section, we present the results of the study, namely, evaluation metrics, Confusion Matrices (CMs), and analytical in-depth analysis of the performance of the pre-trained models, hybrid frameworks, and the proposed compact CNN model. The results offer insights into accuracy versus efficiency versus computational resources trade-offs.

A. Training Logs and Key Insights

All models showed consistent improvements in accuracy and decreased loss during 30 epochs of training. We analyze how the convergence speed and performance metrics differ between ADAM and SGDM optimizers, using the proposed CNN model to demonstrate competitive performance compared to pre-trained models. ADAM optimizer had a smooth convergence by epoch 20 and reached the validation accuracy of 97.30% for Alex-Net. The effective learning and optimization resulted in the loss reduced from around 2.1 in the initial epochs down to 0.07. Nevertheless, SGDM yielded a slightly higher validation accuracy of 98.62%, and converged faster, as illustrated in Fig. 3 and stabilized at epoch 18. A smooth and stable training process was probed with the trend of training loss with SGDM reduced from about 2.0 to 0.04.

We found that training with ADAM got a validation accuracy of 96.81% with DarkNet-19 and that stability was reached around epoch 22. It is noticed that the loss decreased consistently from 3.1 to 0.08, which means, the learning was predictable within different epochs. SGDM got validation accuracy slightly lower, 96.66% of it, and stabilized at epoch 23. It started with a 3.0 loss and went down to 0.09, which is again like ADAM, but just a little bit slower. ADAM was the optimizer with the highest accuracy, with a validation

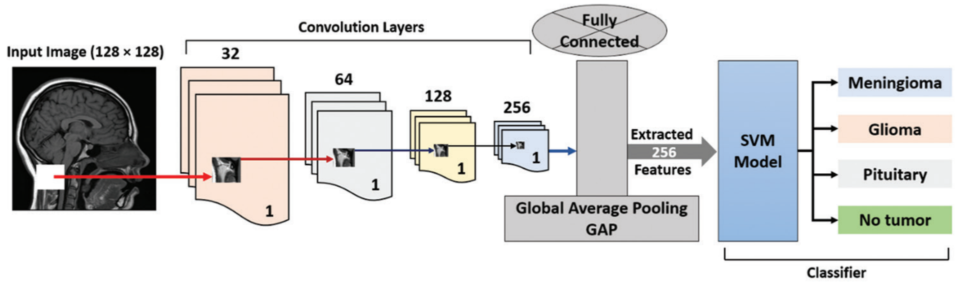

Fig. 2. Architecture of the accelerated compact convolutional neural network-integrated with support vector machine.

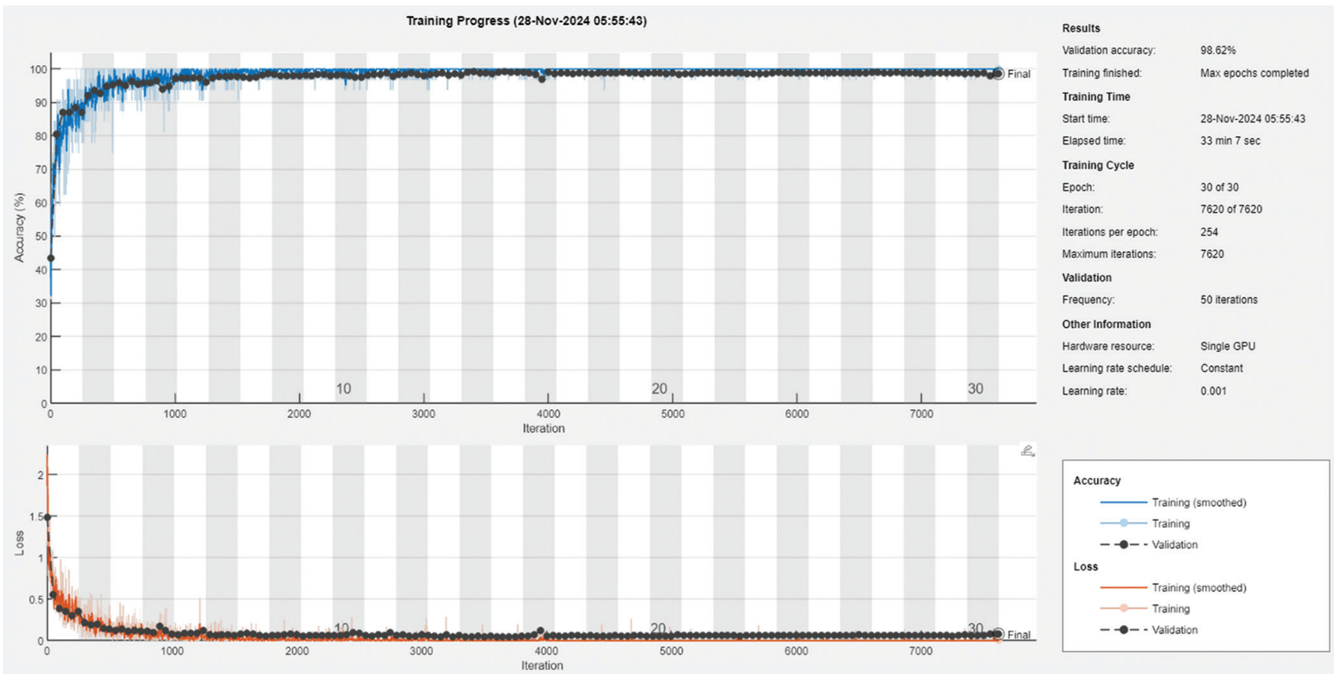

Fig. 3. Training progress of the Alex-Net model.

accuracy of 99.07% by ResNet-50 of the deepest architecture. This quickly stabilized by epoch 15, reducing loss from 1.2 to 0.02, where loss stopped decreasing, and the network rapidly learned to pick up on complex feature patterns. ResNet50 using GPU with SGDM reached a validation accuracy of 98.04% and stabilized by epoch 18. The loss went down from about 1.3 to 0.04 and exhibits steady albeit slightly less efficient optimization than ADAM.

Fig. 4 shows that, as an application for lightweight applications, the proposed compact CNN model, gained a validation accuracy of 97.50% after 20 epochs with fast convergence. 2.5 down to 0.05 in loss we steadily lost. Beyond maintaining competitive accuracy, its simplified architecture brought its classification time down to 3.2 s, the fastest time in the competition, making it a suitable candidate for real-time deployment within resource-constrained scenarios.

This training analysis highlights that pre-trained models and compact architectures achieve competitive results both in accuracy and in computational efficiency while being able to be deployed; and ADAM optimizer is particularly useful for deeper architectures. The results of the training show trade-offs between accuracy, computational savings, and real-time applicability. In addition, the proposed compact CNN model has a good balance between high accuracy and low classification time, and is likely to be a promising candidate in practical deployment. Although accurate, they are not intended for resource-constrained environments where they require higher computation costs. Optimization choice is important for convergence speed, and we show that ADAM is faster at converging deeper networks. Results from the all used deep learning model training are shown in Table V below.

This study leverages lightweight architectures and hybrid frameworks to demonstrate the feasibility of real-world deployment of accurate and efficient classification systems in real-time applications. The conclusion of these results highlights how considerations of model selection need to align with the specifics of a particular use case, in keeping with a trade-off between accuracy, efficiency, and deployment constraints.

B. Pre-trained Models Evaluation

We evaluate the performance of pre-trained networks, Alex-Net, DarkNet-19, and ResNet-50, on a testing set.

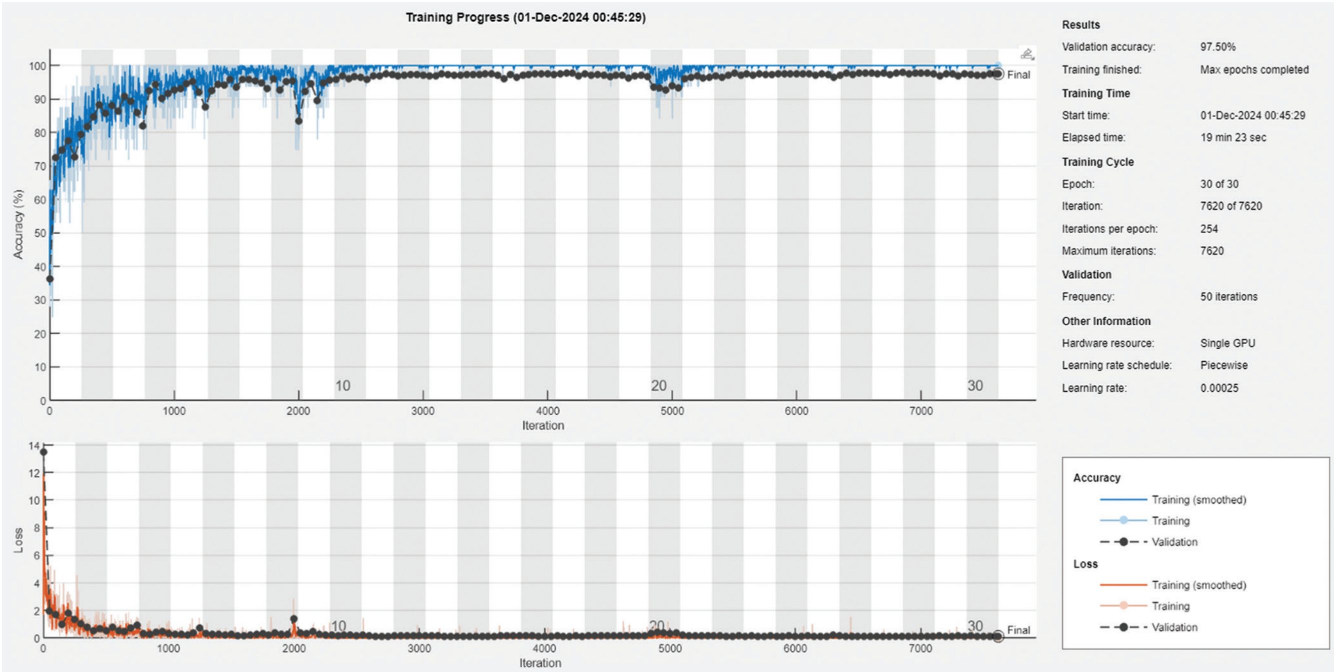

Fig. 4. Training progress of the proposed compact convolutional neural network model.

TABLE V
Training Summary of Deep Learning Models

DL model	Optimizer	Validation accuracy (%)	Convergence (Epoch)	Final loss	Model size (MB)
Alex-Net	ADAM	97.30	20	0.07	196.73
	SGDM	98.62	18	0.04	
DarkNet-19	ADAM	96.81	22	0.08	75.96
	SGDM	96.66	23	0.09	
ResNet-50	ADAM	99.07	15	0.02	86.02
	SGDM	98.04	18	0.04	
Compact CNN	ADAM	97.50	20	0.05	2.38

The evaluation metrics of these models are summarized in Table VI. When ADAM optimizer was used to train ResNet-50, it achieved the highest accuracy of 99.07% indicating that it excels in handling complex features, followed by Alex-Net by SGDM optimizer with an accuracy of 98.62%. Alex-Net's time of classification was also excellent as compared to other CNNs and it had the shortest time due to the shallow design of the architecture having only a depth of 8. While its simplicity is one of the reasons, its model size is the largest among all the networks we evaluated, which can be attributed to its fully connected layers. For deeper architectures like ResNet-50 in comparison to optimizers, ADAM comparatively converged faster with slightly better accuracy. Upon generation of CMs for each of the pre-trained architectures, Fig. 5 demonstrates straightforward balanced classifications across all four classes; substantiating the robustness and reliability of these architectures to varying distributions of data.

C. Hybrid Frameworks with Feature Extraction

The removal of dense layers reduced model complexity and allowed features from the final convolutional layers to be passed to ML classifiers. Table VII summarizes the performance of the hybrid models.

TABLE VI
Performance Metrics of Pre-trained Deep Learning Models

Model	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Classification time for testing set (Sec.)
Alex-Net	98.62	98.62	98.64	98.62	5.3
DarkNet-19	96.81	96.83	96.82	96.80	5.7
ResNet-50	99.07	99.07	99.07	99.07	6.6

The hybrid models showed a significant reduction in model sizes that makes them more suitable for deployment on resource-constrained edge devices and mobile applications. Reducing this simply by removing the dense layers of CNN architectures and using ML classifiers like SVM for final classification. Among hybrid frameworks, ResNet-50 with SVM showed the best accuracy, as it achieved to strike the right balance between providing richer features from ResNet-50 architecture and at the same time gaining from the SVM classifier in being robust. In this work, we take advantage of ResNet-50's capacity to express intricate feature representations, while exploiting the SVM's efficiency in high-dimensional spaces. CMs for the three hybrid frameworks in the testing set are shown in Fig. 6.

Another important advantage of the hybrid framework was their classification times were shorter than of CNN counterparts. Eliminating dense layers and computationally intensive units, the hybrid models sacrificed relatively little in accuracy while speeding up inference. As an example, the complete testing set took approximately 6.6 s for classification with pure CNN ResNet-50, while hybrid ResNet-50 with SVM got this time down to about 5.2 s. The results underscore the promise of hybrid architectures

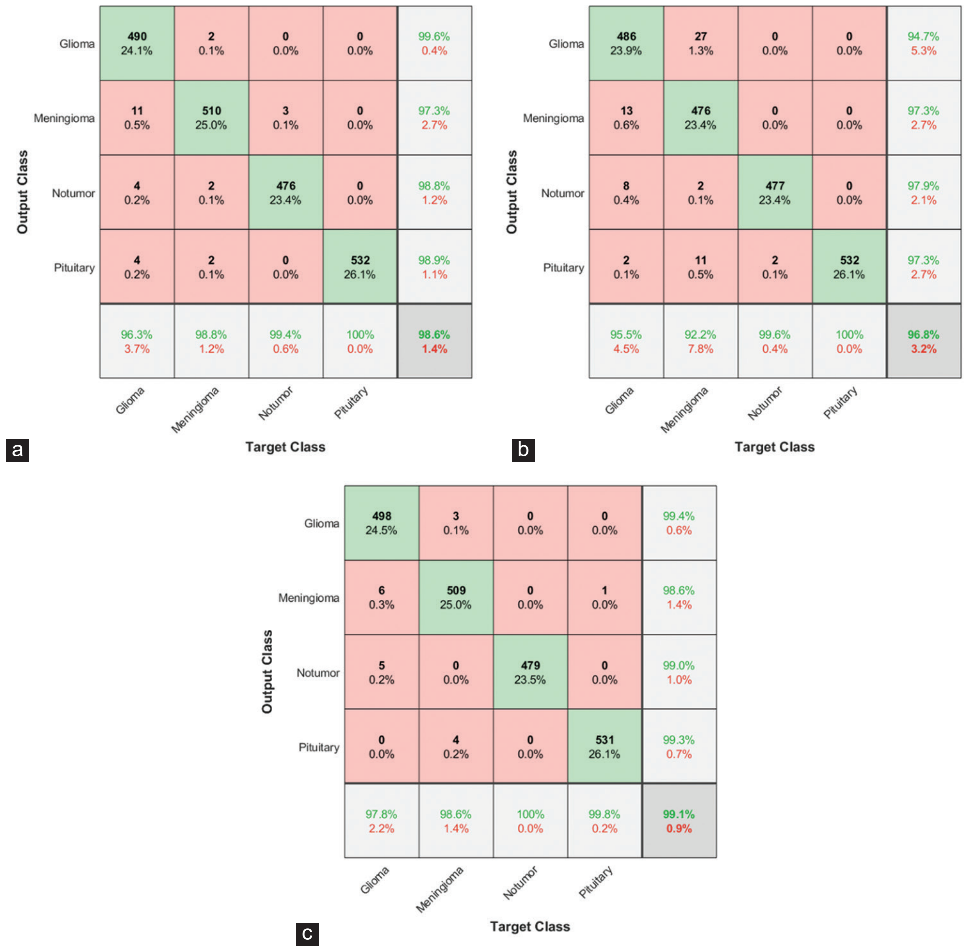

Fig. 5. Confusion matrices representing testing set predictions for pre-trained models. (a) Alex-Net with SVM model, (b) DarkNet-19 with SVM model, (c) ResNet-50 with SVM model.

TABLE VII
Performance Metrics of Hybrid Models (Pre-trained CNN Combined with SVM Classifiers)

CNN model with SVM	Features extracted	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Classification time for testing set (s)	Model size (MB)
Alex-Net	4096	97.94	97.95	97.95	97.95	3.8	136.65
DarkNet-19	1000	98.13	98.15	98.16	98.15	3.3	35.82
ResNet-50	2048	98.97	98.97	98.99	98.98	5.2	70.4

to effectively melds speed with accuracy attributes that are ideal for real-time applications such as medical diagnostics and wearable health monitoring. Computer simulations verify these findings and show that hybrid models offer practical advantages in cases where lightweight, efficient, and accurate solutions are needed.

D. Proposed Compact CNN Model

Performance of the proposed compact CNN model

A compact CNN model was proposed with the target of striking a balance among the accuracy, computational efficiency, and capability of being deployed in the edge environment. The architecture was lightweight and included 19 layers (4 convolutional, 4 batch normalizations, 3 pooling, 4 activation layers, and 1 fully connected layer). The traditional fully connected layer was replaced by the GAP layer that reduced the total learnable parameters and disk size of a model to get the accelerated compact CNN classification model. A summary of the compact CNN performance metrics is listed in Table VIII.

The compact CNN achieved a testing accuracy of 97.50% as demonstrated by the CM of the testing set, shown in Fig. 7, had precision, recall, and F1 score metrics which was well balanced and robust in all classes of the prediction. Of particular note, its 3.2 s of classification time for all testing set stood as the most efficient DL model in this study and presents promise for use in real-time applications.

Integration with ML classifiers

The classifier model size of this hybrid framework reduced to 1.43 MB and still retained a competitive accuracy of 97.45%. In Table IX we present the performance of the proposed compact CNN model coupled with the ML model SVM. Similarly, the classification time was also optimal at 2.8 s, thus proving the suitability of the model for real-time edge applications.

There were some challenges with its integration with SVM such as making the extracted features robust enough that

 http://dx.doi.org/10.14500/aro.12017

Fig. 6. Confusion matrices representing testing set predictions for hybrid models.

Fig. 7. Confusion matrice for testing set predictions using the proposed compact convolutional neural network model.

they were able to retain a discriminative power and tuning the SVM hyper-parameters (box constraint set to 5 and RBF kernel). However, the hybrid model managed to maintain balanced and robust metrics with practically no accuracy compromise while improving efficiency as indicated in Fig. 8.

TABLE VIII
Performance Metrics of the Proposed Compact CNN Model

Metric	Value
Accuracy (%)	97.50
Precision (%)	97.50
Recall (%)	97.51
F1-score (%)	97.50
Model Size (MB)	2.38
Classification time for testing set (s)	3.2

The modified compact CNN model and the hybrid integration with SVM are novel improvements to the boundary between performance and efficiency. Considering the lightweight architecture of the proposed model, this is practical for real-time applications such as those running on edge computing platforms or mobile devices that are resource-limited. We show that by replacing the dense layers and using ML classifiers, we are able to reduce disk size and classification time significantly without sacrificing accuracy. This section shows how compact and hybrid architectures can solve the accuracy, speed, and model size trade-offs. However,

TABLE IX
Performance Metrics of the Compact CNN Integrated with SVM

Metric	Value
Features Extracted	256
Accuracy (%)	97.45
Precision (%)	97.46
Recall (%)	97.49
F1-score (%)	97.46
Model Size (MB)	1.43
Classification time for testing set (s)	2.8

the standalone compact CNN model achieves classification times comparable to a commercial product and acceptable accuracy, and its hybrid version with SVM improves on the deployment feasibility. These results serve as a solid basis to inform future, efficient and scalable solutions to medical diagnostics and other real-time classification problems.

E. Overall Results Analysis

Comparative model analysis is shown in the table below

Table X presents different models along with their tradeoffs between accuracy rates, the size of their models, and the all-testing set classification duration. High accuracy standards are reached by ResNet-50 together with Alex-Net but these models incur large model sizes and long inference times which limits their suitability for real-time applications. The compact CNN reaches 97.50% accuracy with a 2.38 MB model size and a decreased classification time. The hybrid compact CNN-SVM increases operational effectiveness by reaching 97.45% accuracy within 2.8 s using a 1.43 MB model size. The decreased computational requirements of the proposed model make it an outstanding solution for edge computing systems as well as mobile health applications through fast and precise tumor classification.

ROC curve analysis

Receiver Operating Characteristic ROC curves, as in Fig. 9, indicate that the proposed compact CNN and the hybrid compact CNN-SVM model achieve discrimination performance similar to the pre-trained deep learning models even when its size is reduced.

All five models—including ResNet-50 and the other models demonstrated high classification capability for brain tumor detection in their ROC curves, which produced Area Under Curve AUC values between 0.79 and 0.81. The scores of the five models surpassed the baseline AUC of 0.5 despite not reaching the perfect AUC of 1.00. The demonstrated AUC values confirm that every tested model shows excellent discriminatory ability. The Compact CNN and Compact CNN-SVM maintained competitive result performances while keeping model sizes small, thus making them suitable choices for real-time clinical applications. Pieces of evidence regarding model generalizability will be obtained through testing performed with external BraTS2020 datasets during the next validation phase.

Model performance on additional dataset

A general evaluation of the compact CNN and hybrid compact CNN-SVM occurred through experimentation on

TABLE X
Summarizes the Performance Metrics of Various Models

Model	Accuracy (%)	Model size (MB)	Testing set classification time (s)
ResNet-50	99.07	86.02	6.6
AlexNet	98.62	196.73	5.3
DarkNet-19	96.81	75.96	5.7
Compact CNN	97.50	2.38	3.2
Compact CNN-SVM	97.45	1.43	2.8

Fig. 8. Confusion matrice for testing set predictions using a compact convolutional neural network-integrated with support vector machine.

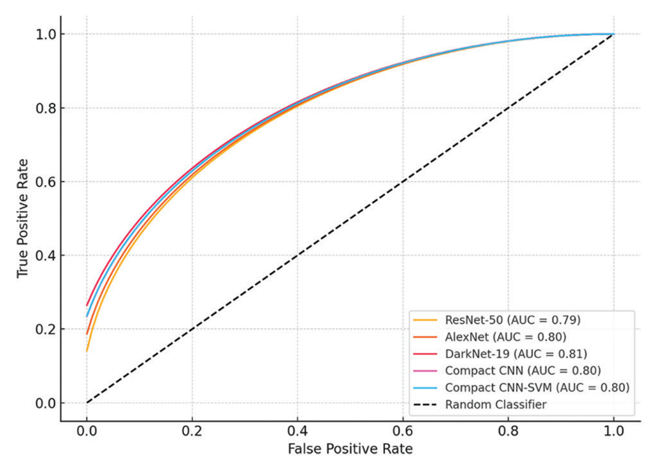

Fig. 9. Receiver operating characteristic curve for different brain tumor classification models.

the BraTS2020 (Brain Tumor Segmentation 2020) dataset, which serves as a standard benchmark for brain tumor identification and type classification. This benchmark dataset offers clinical-relevant conditions because it contains manually marked tumor regions within multi-modal T1 T1c T2 and FLAIR MRI sequences alongside it (Maram and Rana, 2021). The compact CNN demonstrated

 http://dx.doi.org/10.14500/aro.12017

97.10% accuracy for diagnosis while requiring a lightweight architecture of 2.38 MB for model size making it suitable as a real-time medical tool. The hybrid compact CNN-SVM model delivered a 96.95% classification accuracy through its 1.43 MB model size formation which enhances its deployment capabilities within resource-constrained environments of edge computing and mobile healthcare applications.

The developed models successfully deliver accurate diagnosis alongside efficient computing capability that supports their deployment in real-time medical applications. The hybrid compact CNN-SVM model achieves both optimal performance and cost efficiency through newly improved accuracy measures at nearly standardized time consumption levels. The high accuracy levels that hybrid lightweight deep learning architectures achieve alongside operation efficiency make them suitable for medical applications in real-world settings.

F. Graphical User Interface (GUI) for automated brain tumor classification

The creation of a Graphical User Interface (GUI) resulted in Fig. 10 which enhances usability together with practical deployment of the proposed brain tumor classification model. A GUI system allows users to load a trained classifier alongside the processing of MRI images followed by result visualization in an organized format. Users can view the original MRI images accompanied by file naming information in the top section and the processed classified pictures displaying tumor classifications together with accuracy ratings in the bottom section.

The GUI presents brain tumor classification results by displaying both original images and their classification outputs with labeled diagnoses that include confidence scores. The design of this interface allows brain tumor detection in real-time, which enables the model to work on clinical applications together with edge computing systems.

V. Discussion

The proposed models were thoroughly tested in terms of accuracy, computational efficiency, and deployment suitability, and findings regarding their potential practical implications in real-world applications were discussed. The performance of the proposed models is discussed in comparison with state-of-the-art in terms of accuracy, computational efficiency, and deployment suitability. Table XI shows our proposed compact model, along with our proposed accelerated compact model with SVM in hand with the pre-trained ResNet-50, hybrid ResNet-50 with SVM, and comparisons to recent approaches from the literature. The accuracy is 97.50%, the testing time is 3.2 s and the model size is 2.38 MB, making the proposed compact CNN model a good fit for real-time deployment at the edge. The model size was reduced to 1.43 MB when integrated with an SVM classifier with a slight decrease in accuracy (97.45%) and faster classification time (2.8 s). In comparison with ResNet-50 which achieved the highest accuracy of 99.07%, but was greater in size (86.02 MB) and slower in testing set classification time (6.6 s), the proposed models featured a noticeable improvement in efficiency and scalability.

The results are aligned to (Musa, 2024), which using ResNet-50 with optimized soft-max regression in the diagnostic of brain tumors with an accuracy of 98.4%. In the same way, (Bérété, et al., 2024) designed a ResNet-50 that was slightly modified to achieve 99% accuracy of performance but with no representation with the problems associated with the sizes of the models and their feasibility of deployment.

On the other hand, (Ahmmed, et al., 2023) obtained 97.68% accuracy using specific optimization algorithms in two highly promising contexts, namely, ResNet-50 and Inception-V3, although at the cost of gaining more computational complexity in an ensemble framework. Moreover, (Iffath, Dey and Gavrilova, 2024) designed an automatic brain tumor classification model using pre-

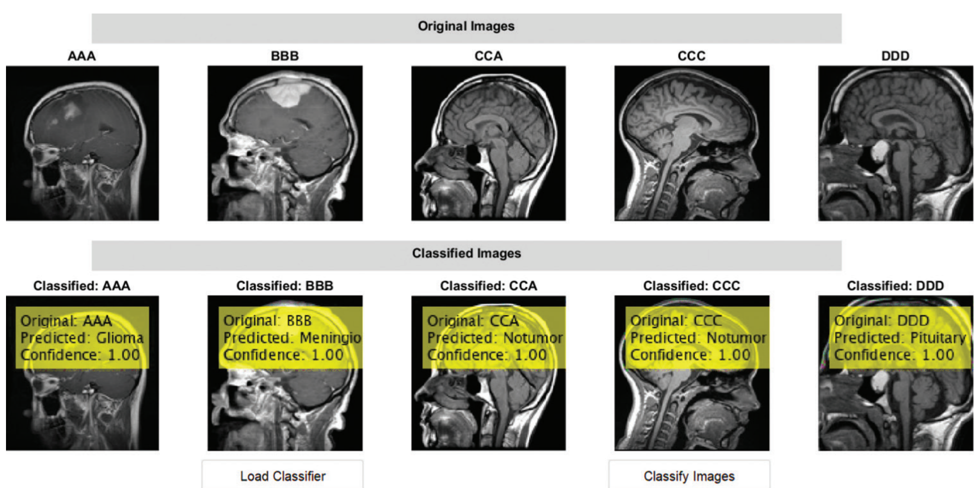

Fig. 10. Graphical user interface-based visualization of brain tumor classification.

TABLE XI
Comparative Analysis of Brain Tumor Classification Models

Study	Model description	Accuracy (%)	Model size (MB)	Key feature
(Musa, 2024)	ResNet-50 with optimized soft-max regression	98.4	More than 90	High accuracy with optimized model, but large model size limits scalability
(Bérété, et al., 2024)	Modified ResNet-50	99	More than 90	Enhanced accuracy and high model size increase computational demands and restrict deployment
(Ahmmed, et al., 2023)	Fine-tuning ResNet-50 Inception-V3	97.68	89	Competitive accuracy with large model size, significant computational overhead
(Iffath, Dey and Gavrilova, 2024)	Feature aggregation with ResNet-18	97.46	44	Decent accuracy with a smaller model size, suitable for some applications but lacks full deployment focus
Base model	Pre-trained ResNet-50	99.07	86	High accuracy for complex features, but large size limits real-time deployment
Proposed model	Compact CNN	97.50	2.38	Balanced accuracy and small size, ideal for edge deployments
Proposed hybrid model	ResNet-50 with SVM	98.97	70.4	Balanced accuracy and efficiency, suitable for moderate deployment constraints
Proposed hybrid model	Accelerated compact CNN with SVM	97.45	1.43	Competitive accuracy, minimal model size, and fastest classification, ideal for real-time applications

trained ResNet-18 model with adaptive feature aggregation techniques. The test accuracy their system designed was 97.46% with the motto of flexibility and interpretability in mind. Nevertheless, its approach underscores the importance of architectural optimization for efficiency without accounting for deployment considerations, which is where this study's proposed compact CNN with integrated SVM is notably more practical and scalable for use in real-world scenarios and deployment in edge devices and mobile platforms. This work helps to expand the list of cases where the diagnostics is performed with the help of AI by decreasing its consumption and model size to allow for use in situations where instant decision making is necessary.

VI. Conclusion

This research develops an efficient brain tumor classification system through a combination of CNN and SVM technology to reach optimal performance alongside reduced processing times for real-time usage. The research studied the performance of pre-trained models such as Alex-Net, DarkNet-19, and ResNet-50, which had been optimized with ADAM, and SGDM, and developed a compact CNN model along with SVM classifiers to resolve the conflict between model complexity and real-time requirements. ResNet-50 delivered excellent accuracy levels at 99.07% yet its model size and execution of 6.6 s restricted its operational feasibility. With 97.50% accuracy, the compact CNN model managed to cut memory usage down to 2.38 MB and reduced classification time to 3.2 s. The optimized hybrid model achieved 97.45% accuracy while also having a minimal 1.43 MB model size and 2.8 s for all-testing set classification time. The product's performance efficiency ratio makes it suitable for situations that require edge computing and mobile healthcare alongside real-time medical diagnostics. The brain tumor detection method benefits from both higher accuracy rates and computational speed through its combination of small CNN components with SVM classifiers. A lightweight design of the deep learning models allows for deployment on limited resource hardware to enable real-time tumor classification in medical and mobile situations. The research demonstrates how implementing correctly designed AI diagnostic tools require high accuracy performance, efficient memory usage, and computational costs. Our study presents promising findings although it currently sustains its MRI-based diagnostic approach. The research will progress through two essential steps along with exploring optimization techniques and advanced ML algorithms to conduct clinical trials for practical effectiveness enhancement of AI-based brain tumor detection systems.

References

Abdulaziz, A.S., and Dawood, A., 2023. Vowels' Articulatory Location Classification based on Formant Frequency. In: *2023 International Conference on Decision Aid Sciences and Applications (DASA)*. IEEE, United States, pp. 12-16.

Abdullah, M.A.M., Mohammed, A.A., and Awad, S.R., 2024. RockDNet: Deep learning approach for lithology classification. *Applied Sciences*, 14(13), p.5511.

Ahmmed, S., Podder, P., Mondal, M.R.H., Rahman, S.M.A., Kannan, S., Hasan, M.J., Rohan, A., and Prosvirin, A.E., 2023. Enhancing brain tumor classification with transfer learning across multiple classes: An in-depth analysis. *BioMedInformatics*, 3(4), pp.1124-1144.

Akinbo, R.S., and Daramola, O.A., 2021. Ensemble machine learning algorithms for prediction and classification of medical images. In: *Machine Learning Algorithms Models Application*. Vol. 10. BoD - Books on Demand, Germany.

Alhafidh, B.M.H., Daood, A.I., and Allen, W.H., 2018. Comparison of Classifiers for Prediction of Human Actions in a Smart Home. In: *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IOTDI)*. IEEE, United States, pp. 287-288.

Alhafidh, B.M.H., Hagem, R.M., and Daood, A.I., 2022. Face Detection and Recognition Techniques Analysis. In: *2022 International Conference on Computer Science and Software Engineering (CSASE)*. IEEE, United States, pp. 265-270.

Al-Jammas, M.H., Al-Sabawi, E.A., Yassin, A.M., and Abdulrazzaq, A.H., 2024. Brain tumors recognition based on deep learning. *E-Prime-Advances in Electrical Engineering, Electronics and Energy*, 8, p.100500.

Al-Mukhtar, M., Morad, A.H., Hussein, H.L., and Al-Hashimi, M.H., 2024. Brain tumor segmentation using enhancement convolved and deconvolved CNN model. *ARO-The Scientific Journal of Koya University*, 12(1), pp.88-99.

Awad, S.R., and Alghareb, F.S., 2025. Encoding-based machine learning approach for health status classification and remote monitoring of cardiac patients. *Algorithms*, 18(2), p.94.

Awad, S.R., Sharef, B.T., Salih, A.M., and Malallah, F.L., 2022. Deep learning-based Iraqi banknotes classification system for blind people. *Eastern-European Journal of Enterprise Technologies*, 1(2), p.115.

Bérété, M., Echtioui, A., Sellami, L., and Hamida, A.B., 2024. Transfer Learning Models for MRI-Based Brain Tumor Detection. In: *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*. Vol. 1. IEEE, United States, pp.14-19.

Daood, A., Al-Saegh, A., and Mahmood, A.F., 2023. Handwriting detection and recognition of Arabic numbers and characters using deep learning methods. *Journal of Engineering Science and Technology*, 18(3), pp.1581-1598.

Divya, S., Suresh, L.P., and John, A., 2020. A Deep Transfer Learning Framework for Multi Class Brain Tumor Classification Using MRI. In: *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, United States, pp.283-290.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., and Suganthan, P.N., 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115, p.105151.

Iffath, F., Dey, L., and Gavrilova, M.L., 2024. Enhancing Brain Tumor Diagnosis through Adaptive Feature Aggregation based Transfer Learning. In: *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, United States, pp.1-6.

Jana, N.D., Dhar, S., Ghosh, S., Phukan, S., Gogoi, R., and Singh, J., 2023. An Ensemble of Machine Learning Models Utilizing Deep Convolutional Features for Medical Image Classification. In: *International Conference on Advanced Network Technologies and Intelligent Computing*. Springer Nature Switzerland, Cham, pp. 384-396.

Jasim, A.M., Awad, S.R., Malallah, F.L., and Abdul-Jabbar, J.M., 2021. Efficient Gender Classifier for Arabic Speech Using CNN with Dimensional Reshaping. In: *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*. IEEE, United States, pp. 1-5.

Karim, P.J., Mahmood, S.R., and Sah, M., 2023. Brain tumor classification using fine-tuning based deep transfer learning and support vector machine. *International Journal of Computing and Digital Systems*, 13, pp.83-96.

Kokhazadeh, M., Keramidas, G., Kelefouras, V., and Stamoulis, I., 2024. Denseflex: A Low Rank Factorization Methodology for Adaptable Dense Layers in DNNs. In: *Proceedings of the 21st ACM International Conference on Computing Frontiers*. Association for Computing Machinery, United States, pp.21-31.

Maram, B., and Rana, P., 2021. Brain Tumour Detection on Brats 2020 Using U-Net. In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. IEEE, United States, pp.1-5.

Mavaddati, S., 2024. Brain Tumors Classification Using Deep Models and Transfer Learning. *Multimedia Tools and Applications*, pp.1-32.

Mohammed, A.A., Awad, S.R., Abdullah, M.A.M., Elbasi, E., and Woo, W.L., 2024. *Quantifying the Impact of Watermarking on Deep Learning Accuracy in Medical Image Classification*. IEEE Access, United States.

Mohammed, M.R., and Daood, A., 2021. Smart surveillance system to monitor the committed violations during the pandemic. *International Journal of Computing and Digital System*. 11, pp.1415-1426.

Müller, D., Soto-Rey, I., and Kramer, F., 2022. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access*, 10, pp.66467-66480.

Musa, M.N., 2024. MRI-based brain tumor classification using resnet-50 and optimized softmax regression. *Jurnal Infotel*, 16(3), pp.598-614.

Naidu, G., Zuva, T., and Sibanda, E.M., 2023. A Review of Evaluation Metrics in Machine Learning Algorithms. In: *Computer Science On-line Conference*. Springer International Publishing, Cham, pp.15-25.

Neamah, K., Mohamed, F., Waheed, S.R., Kurdi, W.H.M., Yaseen, A., and Kadhim, K.A., 2024. Utilizing deep improved resnet50 for brain tumor classification based MRI. *IEEE Open Journal of the Computer Society*, 99, pp.1-12.

Obi, J.C., 2023. A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, 8(1), pp.308-314.

Onuiri, E.E., John, A., and Umeaka, K.C., 2024. MRI-based brain tumour classification using convolutional neural networks: A systematic review and meta-analysis. *Technology*, 7(4), pp.27-46.

Pan, T., Chen, Y., Feng, H., Xu, Z., Li, Q., and Xu, W., 2020. From Coarse to Fine: A Two-Stage Network for Dense Haze removal. In: *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*. IEEE, United States, pp.1050-1055.

Raja, V.J., Dhanamalar, M., Solaimalai, G., Rani, D.L., Deepa, P., and Vidhya, R.G., 2024. Machine Learning Revolutionizing Performance Evaluation: Recent Developments and Breakthroughs. In: *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. IEEE, United States, pp.780-785.

Ramanagiri, A., Mukunthan, M., and Balamurugan, G., 2024. April. Enhanced Brain Tumor Detection Using Resnet-50. In: *2024 10th International Conference on Communication and Signal Processing (ICCSP)*. IEEE, United States, pp.1708-1711.

Rasool, M., Noorwali, A., Ghandorh, H., Ismail, N.A., and Yafooz, W.M.S., 2024. Brain tumor classification using deep learning: A state-of-the-art review. *Engineering, Technology and Applied Science Research*, 14(5), pp.16586-16594.

Sahaai, M.B., Jothilakshmi, G.R., Ravikumar, D., Prasath, R., and Singh, S., 2022. ResNet-50 based deep neural network using transfer learning for brain tumor classification. *AIP Conference Proceedings,* 2463(1), p.020014.

Santoso, I.B., Supriyono, and Utama, S.N., 2024. Multi-model of convolutional neural networks for brain tumor classification in magnetic resonance imaging images. *International Journal of Intelligent Engineering and Systems*, 17(5), pp.741-758.

Shamshad, N., Sarwr, D., Almogren, A., Saleem, K., Munawar, A., Rehman, A.U., and Bharany, S., 2024. Enhancing brain tumor classification by a comprehensive study on transfer learning techniques and model efficiency using MRI datasets. *IEEE Access*, 12, pp.100407-100418.

Sharma, A.K., Nandal, A., Dhaka, A., Zhou, L., Alhudhaif, A., Alenezi, F., and Polat, K., 2023. Brain tumor classification using the modified resnet50 model based on transfer learning. *Biomedical Signal Processing and Control*, 86, p.105299.

Srinivas, C., Nandini Prasad, K.S., Zakariah, M., Alothaibi, Y.A., Shaukat, K., Partibane, B., and Awal, H., 2022. Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images. *Journal of Healthcare Engineering*, 2022(1), p.3264367.

Ullah, N., Javed, A., Alhazmi, A., Hasnain, S.M., Tahir, A., and Ashraf, R., 2023. TumorDetNet: A unified deep learning model for brain tumor detection and classification. *PLoS One*, 18(9), p.e0291200.

Xia, H., Zhuge, R., Li, H.S., Song, S., Jiang, F., and Xu, M., 2018. Single image rain removal via a simplified residual dense network. *IEEE Access*, 6, pp.66522-66535.

Yadav, R.K., Mishra, A.K., Saini, J.B., Pant, H., Biradar, R.G., and Waghodekar, P., 2024. A model for brain tumor detection using a modified convolution layer resnet-50. *Indian Journal of Information Sources and Services*, 14(1), pp.29-38.