IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 February 2025, accepted 24 March 2025, date of publication 28 March 2025, date of current version 18 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3555638

RESEARCH ARTICLE

Transforming Brain Tumor Detection Empowering Multi-Class Classification With Vision Transformers and EfficientNetV2

ANEES TARIQ[1], MUHAMMAD MUNWAR IQBAL[1], MUHAMMAD JAVED IQBAL[1], AND IFTIKHAR AHMAD[2], (Member, IEEE)

[1]Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan
[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Muhammad Munwar Iqbal (munwariq@gmail.com)

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under Grant GPIP: 1265-611-2024.

ABSTRACT Brain Tumor is a significant health challenge, with early, precise detection crucial for improving patient outcomes. Traditional diagnostics methods depend on expert radiologists, leading to subjectivity and time-intensive evaluations. In recent advancements, AI and DL have demonstrated remarkable medical imaging potential, yet many existing models are limited to binary classification, failing to distinguish between multiple tumor types. This limitation restricts their clinical applicability and effectiveness in comprehensive diagnosis. We propose a deep learning-based multi-class classification approach for brain tumor detection, leveraging EfficientNetV2 and Vision Transformer (ViT) models to address this challenge. EfficientNetV2, known for its optimized convolutional architecture, achieves high accuracy while maintaining computational efficiency, whereas ViT, a transformer-based model, effectively captures global contextual information in medical images. Our approach was evaluated on a brain tumor MRI dataset containing 7,023 images, divided into training and testing sets. EfficientNetV2 achieved an accuracy of 95% with a loss of 0.13, alongside an F1-score, precision, and recall of 0.96. In comparison, ViT attained 90% accuracy with a loss of 0.30 and an F1-score, precision, and recall of 0.89. The highest accuracy 96%, was achieved by the proposed model using the geometric mean ensemble learning technique. The results demonstrate that our proposed system outperforms many existing deep learning models, offering a robust solution for multi-class brain tumor classification. Our approach enhances diagnostic accuracy by integrating convolutional and transformer-based architectures, aiding radiologists in early and automated tumor detection. This research contributes to the advancement of AI-driven medical diagnostics and has the potential to improve clinical decision-making in brain tumor assessment.

INDEX TERMS EfficientNetV2, convolutional neural network (CNN), brain tumor vision transformers (ViT), magnetic resonance imaging (MRI), multilayer perceptron (MLP).

I. INTRODUCTION

Brain tumors, comprising a diverse array of approximately 200 aberrant tissue growths, present a formidable challenge to human health. Gliomas, originating in the glial cell, are the most common type of malignant BT, comprising 80% of such cases [1]. The World Health Organization classifies gliomas into four categories, each encompassing a spectrum of malignancies ranging from benign to severe. This intricate grading process, essential for determining the appropriate course of treatment, demands precision and is often time-consuming. Magnetic resonance imaging is an essential diagnostic because it is non-invasive and gives precise insights into brain architecture. The intricacy of the grading system and brain biopsies, however, make it challenging to identify brain tumors promptly and accurately. The presence

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

of other brain tumor subtypes, such as meningiomas and pituitary tumors, further exacerbates the clinical intricacy of these patients. Because they originate from the pituitary gland, which oversees regulating hormones, pituitary tumors can either be benign and cause bone development, or they can be malignant and cause long-term hormone imbalances and vision problems. The significance of early detection and precise classification of these tumors cannot be overstated, as delays in diagnosis can significantly impact patient survival rates. While biopsies and manual grading by experts remain the gold standard, the intricate nature of brain tumors and the associated time constraints necessitate innovative and efficient solutions. Magnetic Resonance Imaging (MRI) has proven indispensable in this regard, yet the need for expert interpretation and the intricate grading process constrains its potential.

Xu et al. [2] developed an inertial microfluidic sorting device that has a 3D-stacked multistage intended for circulating tumor cells (CTCs) with efficient downstream analysis. The initial step includes a trapezoidal cross-section for maximum separation, followed by symmetrical square serpentine channels for more enrichment. This novel design produces a particle separation of 92.37% efficiency and 98.10% purity while separating tumor cells from red blood cells at high flow rates yields over 80% efficiency and over 90% purity. This approach allows for quick and integrated CTC analysis.

Recent advancements in vision models, such as Vision Language Models (VLLMs) and other emerging technologies, have the potential to significantly enhance the performance of brain tumor detection methods. VLLMs, for instance, combine vision and language processing, allowing for a more nuanced understanding of both visual data and associated medical narratives. Incorporating these models could improve our system's ability to interpret complex MRI scans in the context of clinical documentation, patient history, or radiology reports. Furthermore, other emerging technologies, such as self-supervised learning and attention mechanisms, are showing promise in enhancing the ability of models to learn from limited data, which would be particularly beneficial for improving performance on rare tumor types. We will consider integrating these advancements in future work to further enhance the robustness and accuracy of our model.

ML, particularly the advent of the DL techniques, offers promising avenues for addressing these challenges. Convolutional Neural Networks (CNNs) and independent learning methods such as autoencoders have demonstrated utility in medical image analysis, aiding in disease prediction, image classification, and tissue segmentation [3]. There is a need for improved frameworks since current research on deep-learning-based brain tumor diagnostics shows inconsistent model performance across different datasets. This study presents a useful solution that makes use of a vision transformer-based methodology to address these issues. The study aims to classify brain images automatically and accurately into four classes: pituitary, glioma, malignant glioma, and no tumor. Vision transformers allow for effective feature extraction, facilitating the identification of crucial object features within these complex images [4]. By contributing to refining existing frameworks, this study tries to bridge the performance gaps observed in deep learning models across diverse datasets. The goal is to revolutionize brain tumor early detection and categorization by developing a strong computer-aided diagnosis tool. This would decrease the need for labor-consuming human treatments and increase the efficacy of treatment plans. This research paper has following contributions:

- The proposed model, with the hierarchical representations and the attention process, vision transformers can capture complex patterns and features found in medical scans, result, the various forms of brain tumors can be identified and classified with greater accuracy and reliability.
- It has scalability and generalization as essential features of vision transformers for managing large-scale datasets and taking into account various variants of brain tumors.
- The use of vision transformers improves the accuracy of the multi-class brain tumor detection technique.
- It presents the robust performance across various scanner types and imaging techniques is ensured by their capacity to learn representations from vast amounts of data and capture global context. This facilitates broader acceptance and possible improvements in patient care by enabling researchers and clinicians to adapt the established models to a variety of clinical contexts.

The rest of the paper is organized as follows: Section II comprehensively presents related work and a literature review. Our proposed methodology is described in detail in Section III, and the results of our analysis are presented in Section IV. Finally, we conclude with a discussion of our findings and suggestions for future work in Section V.

II. LITERATURE REVIEW

Ahmmed et al. [5] had a primary focus on the utilization of data-driven techniques, specifically employing deep learning models such as ResNet 50 and Inception V3, to diagnose BT based on MRI images. This research makes a substantial contribution by meticulously curating paired datasets, each possessing distinct characteristics. It integrates crucial techniques like Early Stopping and ReduceLROnPlateau for hyperparameter optimization. The implementation of strategic innovation, such as tailored pooling and regularization layers, results in high classification accuracy. In this dataset, ResNet 50 combined with Nadam performed exceptionally well, with 99.34% accuracy rates for gliomas, 93.52%, 98.68%, and 97.70% for other class tumors. It is used to test the model at a very small number of test images, which is 692, which is why it achieves high accuracy. ResNet50 with the Adam optimizer shines in a two-class dataset, achieving 97.84% of overall accuracy.

In recent advancements in brain tumor classification, attention-based methods have proven effective. The ConvAttenMixer, presented in [6], is a model based on transformers that integrate convolutional layers, self-attention, and external attention mechanisms to enhance the analysis of MRI brain images. By incorporating two convolution mixer blocks, the model effectively captures the data's spatial and channel-wise dependencies. Self-attention is employed to prioritize local features, while external attention focuses on capturing global interactions. The classification head of the model incorporates a squeeze-and-excitation mechanism. In performance evaluation, the ConvAttenMixer surpasses baseline models, achieving an impressive accuracy of 0.9794 on 5712 MRI scan images. This accuracy outperforms baseline models, which range from 0.87 to 0.93. Notably, ConvAttenMixer exhibits superior precision, recall, and f-measure, highlighting its efficacy in processing both local and global features with reduced computational memory requirements.

In the study [7], The substantial influence of artificial intelligence, particularly deep learning, on the advancement of medical image processing and analysis is investigated. The complete examination includes a variety of tasks, such as illness detection, categorization, and anatomical structure segmentation. The overview begins with foundational principles and then moves on to cutting-edge models such as convolutional neural networks, recurrent neural networks, and generative adversarial networks (GANs). Several medical imaging application fields are comprehensively addressed, including neurology, pulmonary imaging, cardiac imaging, breast imaging, digital pathology, and retinal analysis. Through a critical analysis of both strengths and limitations, the review highlights challenges such as the scarcity of annotated data and image variability.

In research by Generexu [8], a deep learning methodology is introduced to accurately detect and classify intracranial hemorrhage (ICH) subtypes in CT scans. This approach addresses the complexities posed by the intricate brain anatomy and the diverse appearances of hemorrhages. The proposed method combines a 2-D convolutional neural network with a bidirectional long-short-term memory (Bi-LSTM) module, integrating a multi-head attention mechanism to enhance performance. Results from experiments conducted on three benchmark datasets, namely RSNA, CQ500, and PhysioNet, showcase the system's high performance and generalizability. Incorporating the multi-head attention mechanism notably reduces the weighted multilabel binary cross-entropy with logit loss score on the RSNA dataset. The CQ500 dataset exhibits competitive results, achieving 0.959, 0.974, 0.958, and 0.977 of accuracy, sensitivity, specificity, and precision. Impressive 0.9869 AUC scores, 0.9797, and 0.9778 are attained on the RSNA, CQ500, and PhysioNet datasets. These outcomes indicate the potential deployment of the proposed model as an intelligent assistive tool for radiologists to efficiently diagnose ICH. Nevertheless, it is crucial to acknowledge challenges associated with real-world clinical scenarios, such as variations in CT scanner parameters and patient demographics.

S. Chen et al. worked on RNA adenosine modifications, which play an important role in epigenetic control in various biological processes, including cancer. However, the expression of these genes and their prognostic significance in osteosarcoma (OS) remain unknown [9]. One of the primary goals of genome research is to anticipate phenotypes and identify key gene biomarkers. However, three major issues arise when analyzing genomics data to predict phenotypes and choose gene markers. It includes big p and small n, limited repeatability of the chosen biomarkers, and significant noise [10].

In this article [11], the authors describe a machine learning approach for detecting tumor tissue origins using gene length-normalized somatic mutation sequencing data. This data was obtained from the International Malignancy Genome Consortium's (ICGC) Data Portal (Release 28), which included 4909 samples from 13 different cancers. The framework uses a Random Forest (RF) technique with 10-fold cross-validation to provide the best results on a 600-gene dataset. The model achieved an average accuracy of 0.8822 and an F1-score of 0.8886 for the 13 cancer types. This work offers an effective computational methodology for finding the primary location of malignancy by combining DNA sequencing and sophisticated machine learning algorithms. The positive results indicate that this technique might significantly improve clinical cancer detection, giving a reliable tool for tracking the origins of diverse tumors using genomic data.

The research [12] focuses on the pressing global health issue of brain tumors, emphasizing the crucial need for prompt and accurate diagnosis through MRI. It recognizes the shortcomings and time-consuming nature of radiologists' manual assessments, and there is a need for the development of computer-aided classification models. However, existing models often face challenges in terms of performance and explainability, leading to skepticism among physicians. The study introduces an innovative classification and localization model, utilizing pre-trained models. The experimental findings reveal that the pre-trained models and other DL techniques for both classification accuracy and visualization results. This approach showcases the potential to reduce diagnostic uncertainty and improve the validation of brain tumor classification.

The emergence of transfer learning in medical image analysis, particularly for 3D medical datasets, addresses the challenge of limited annotated data for training deep learning models. The study [13] introduces a novel approach, the Medical Transformer. Employing a multi-view strategy across three planes of the 3D volume enhances spatial relations while maintaining parameter efficiency during training. The suggested framework involves pre-training the model using self-supervised learning on an extensive dataset of normal

brain MRI scans, with a focus on predicting masked encoding vectors.

Research work [14] presents electromyogram-based control for self-paced operation. This system produces configurable navigation objectives and uses event-related potentials to pick targets, allowing robots to navigate autonomously. Experiments with eight individuals demonstrated that this approach shortens work length, decreases command frequency, and improves navigation patterns, hence improving human-robot integration in unstructured contexts.

Si et al. [15] used functional near-infrared spectroscopy (fNIRS) and deep learning to create a dual-branch joint network (DBJNet) that detects emotions across subjects. Our trials using video stimuli showed that fNIRS can accurately discriminate emotions with high F1 scores. The model obtained 74.8% accuracy for three-category emotion detection and more than 89% accuracy for two-category tasks, highlighting fNIRS's promise for emotion decoding and improving affective brain-computer interfaces.

This work [16] investigated the decoding of moral elevation using EEG data from 23 persons who watched films that elicited this sensation. We achieved good prediction performance using power spectra characteristics and regularized regression analyses ($r = 0.44 \pm 0.11$). Our findings show that EEG may accurately predict moral elevation and that small-sample brain data can reflect ongoing moral elevation experiences as identified by crowdsourced danmaku comments.

Deep neural networks have emerged as crucial tools in medical image classification despite encountering challenges like the vanishing gradient problem and overfitting. In a recent study [17], an approach was introduced employing a deep network model that utilizes ResNet-50 and global average pooling to address these issues. The proposed model underwent simulation using a dataset of 3064 images from three-tumor brain magnetic resonance imaging. The evaluation revealed impressive results, achieving a mean accuracy of 97.08% with data augmentation and 97.48% without data augmentation. These outcomes surpass the performance of existing models in terms of classification accuracy, validating the effectiveness of the proposed model in improving brain tumor detection and classification.

The study [18] presents a methodology to overcome constraints observed in traditional Vision Transformer (ViT) models designed for medical image segmentation. The focus of this approach is to enhance generalizability across various organ regions. The suggested universal ViT segmentation model utilizes task-specific prompts to amalgamate features from the encoder of the ViT-based segmentation model with trainable universal prompts. This innovative framework enhances flexibility and efficiency, allowing for the segmentation of various organ regions within a unified model architecture. Empirical validation across multiple medical image datasets demonstrates its effectiveness and adaptability. The results indicate improvements in segmentation performance, showcasing the potential of this forward-looking methodology for advancing medical image analysis and supporting various healthcare applications.

Since the advent of MRI scans in 1978, researchers at EML Laboratories paved the way for advanced diagnostic capabilities. In 2020 alone, an estimated 251,329 lives were claimed by primary cancerous brain and CNS tumors. The conventional method of relying on histological subtyping during brain MRI interpretation poses subjectivity challenges for radiologists. This issue is further exacerbated in developing countries like India, where the doctor-to-population ratio is 1:1151. To address these challenges, the study [19] employs swin transformers and deep learning techniques to detect, classify, locate, and determine the size of brain tumors in MRI scans. The results showcase a promising approach to enhancing the efficiency of radiologists and doctors, offering a potential breakthrough in early tumor detection and ultimately saving lives.

In the study [20], the researchers tackle the challenge of accurately identifying infiltrative glioma tumor regions within brain tissue, a critical aspect for enhancing prognostic outcomes in neurosurgery. They introduce an innovative approach, the MLS-CNN employs a sigmoid activation function to train a dataset on 59,811 image patches extracted from 73 brain tissue samples. The model achieves impressive metrics with a high accuracy rate of 91.70%, sensitivity of 91.62%, and specificity of 91.78%, outperforming the current state-of-the-art ViT in distinguishing infiltrating edges. Additionally, the MLS-CNN exhibits notable computational efficiency, producing predictions within 11.5 seconds, compared to ViT's 81.4 seconds.

The study [21] focuses on the critical need for quick identification of malignant brain tumors and presents a hybrid CNN architecture that incorporates InceptionV3, ResNet-50, VGG16, and DenseNet for classification. The approach employs a Kaggle dataset of 3929 pictures, including 2556 non-tumorigenic and 1373 tumorigenic cases, to extract features from MRI images, segment tumors using mask images, fuse segmented tumor images with originals, and classify them using four different CNN models. The models are further optimized with one API to enhance performance. The evaluation employs mean Intersection over Union (mIoU) as a key metric for assessing the balance between true positives, false positives, and false negatives. This approach offers a promising avenue for efficient and accurate brain tumor classification, with the optimization revealing insights into the individual model performances.

The study [22] extensively explores the transformative influence of vision transformers on medical image analysis. Vision transformers play a crucial role in elevating the precision of medical image processing, showcasing unparalleled accuracy in identifying anomalies and diseases through meticulous image segmentation. Moreover, their robust capabilities in classification, advanced object detection, image restoration, and synthesis significantly contribute to precise

diagnosis and effective treatment planning. Despite considerable advancements, challenges such as interpretability, model complexity, and the necessity for extensively annotated datasets persist, underscoring the importance of addressing these obstacles for widespread adoption. The study concludes by highlighting promising future directions, envisioning personalized medicine, real-time diagnostics, and collaborative healthcare platforms as potential areas where vision transformers can continue to revolutionize healthcare technology.

The study [23] focuses on brain cancers using MRI. The CNN with Multi-Branch Network with Inception block showed exceptional performance, achieving a fantastic 99.30% of accuracy rate with a variation of only 0.0005. It demonstrates its clear dominance over existing state-of-the-art models in the sector.

In this work, [24] created a semi-supervised stimulated Raman CycleGAN model that can quickly convert fresh-tissue SRS pictures to H&E stains for intraoperative histology. Within 3 minutes, this approach generates stimulated Raman virtual histology (SRVH), which closely matches genuine H&E staining results. SRVH enables board-certified neuropathologists to discriminate histologic subtypes of human glioma more efficiently than traditional SRS pictures, resulting in faster and more accurate intraoperative diagnoses.

Hu et al. [25] present TMPLiTS, a new multi-phase liver tumor segmentation approach based on contrast-enhanced computed tomography (CECT) data. TMPLiTS uses uncertainty estimates and Dempster-Shafer Evidence Theory (DST) to get trustworthy segmentation findings. A multi-expert mixing strategy (MEMS) guarantees a strong fusion of multi-phase pieces of evidence, which improves segmentation accuracy and interpretation. Experimental findings show that TMPLiTS outperforms previous approaches, with robust performance against perturbations.

The study [26] presents a novel approach to malaria diagnosis using transformer models and generative adversarial networks (GANs). The framework optimally balances high accuracy and low resource consumption in the multi-class Plasmodium classification of thin blood smear microscopic images by leveraging transformer advantages like fine-grained feature extraction and attention mechanisms. Robustness is enhanced using a GAN to generate extended training samples. The Swin Transformer achieves superior detection performance with 99.8% accuracy. These findings highlight the potential of transformer models in enhancing malaria diagnosis efficiency.

The study [27] tackles the complexities associated with diagnosing brain and central nervous system cancers by proposing a transfer-learning-based artificial intelligence approach utilizing CNNs for precise and non-invasive grading and classification. The study involved the development of five clinically relevant multi-class datasets, varying from two to six classes. It assessed the performance of the CNN-based model against six traditional machine-learning classification methods. Results consistently demonstrated the superior performance of the CNN-based DL model over the ML models across all five multi-class tumor datasets. Specifically, the AlexNet transfer learning system achieved mean accuracies of 100%, 95.97%, 96.65%, 87.14%, and 93.74% for the two-, three-, four-, five-, and six-class datasets. Additionally, the mean areas under the curve for DL and ML were 0.99 and 0.87, revealing a notable 12.12% improvement in favor of DL. The study concluded that the transfer-learning-based AI system proves to be a valuable tool for multi-class brain tumor grading, exhibiting superior performance in comparison to traditional ML systems.

The Unified Visualization and Classification Network (UniVisNet), presented in the study [28], offers a groundbreaking framework to address the challenges associated with accurate glioma grading. UniVisNet excels in both classification performance and the generation of visual explanations by incorporating a subregion-based attention mechanism and multiscale feature map fusion. The Unified Visualization and Classification head (UniVisHead) simplifies the process by directly producing visual explanations. Through extensive experiments, UniVisNet consistently outperforms baseline models and prevalent visualization methods, delivering remarkable results in glioma grading, including an accuracy of 89.3%, AUC of 94.7%, specificity of 85.3%, and sensitivity of 90.4%. UniVisNet's visually interpretable explanations surpass existing approaches, showcasing its potential for clinical applications in leveraging deep learning for comprehensive insights into glioma spatial heterogeneity.

In recent years, advancements in medical image analysis leveraging deep learning, particularly Generative Adversarial Networks (GANs), have shown promise in augmenting training datasets for improved classification accuracy. However, limitations arise from insufficient clinical data. The study [29] introduces an approach that addresses imbalanced datasets in brain tumor classification using MRI, outperforming baseline models on two datasets. In the BraTS 2019 dataset, the proposed method demonstrated superior performance, surpassing baseline models up to an AUC of 6.4%, F1-score of 3.4%, and accuracy of 5.4%. Improvements were notable on an internal pediatric LGG dataset, with a 4.3% increase in AUC, 7.3% in F1-score, and 9.2% in Accuracy. The findings underscore the effectiveness of utilizing generated tumor ROIs to tackle imbalanced data issues. This approach holds promise for achieving precise diagnoses of uncommon brain tumors using MRI scans.

In recent years, image analysis in medicine has become a focal point for researchers, particularly in assessing the severity of medical conditions and predicting outcomes. However, the accuracy of noise-trimming outcomes has diminished with the increasing complexity of trained images, leading to lower prediction accuracy scores. Addressing this challenge, the study [30] introduces a machine-learning prediction framework for evaluating the severity of brain tumors using MRI scans. The study leveraged the boosting function to

demonstrate improved error-pruning results. The Proposed Solution function is employed for successful feature analysis and tumor prediction operations. Evaluation in the Python environment and comparative analysis reveal that the original MLPM model exhibits the highest precision in predicting brain tumors, marking a significant advancement in the field.

In previous work [31], the transformative effect of transformers in medical image processing has revolutionized their successful applications in segmentation, reconstruction, classification, and diagnosis. The paper begins by elucidating transformers' fundamental principles and model structures, emphasizing key enhancements such as integration with the UNet network, development of lightweight variants, reinforcement of cross-fast link mechanisms, and construction of large models with transformers as the core. The subsequent sections delve into a comprehensive discussion of the various applications, highlighting the significant strides made in medical image analysis. The review concludes by addressing the challenges faced by transformers in the medical imaging domain and presenting future development prospects. This study serves as a valuable reference for researchers exploring the dynamic intersection of transformer technology and medical image processing.

This work [32] proposes a dual-domain diffusion model (DDDM) for sparse-view CT reconstruction that improves picture quality while eliminating artifacts. DDDM is made up of two modules: sinogram upgrading (SUM) and image refining (IRM), which work together to solve deterioration caused by sparse sampling. SUM gradually improves sinograms, reversing CT image degeneration, whereas IRM minimizes artifacts and recovers features. DDDM surpasses deep-learning baseline models in quality criteria and has high generalizability to untrained organs.

This paper [33] suggests two ways to speed up Transformer models for picture feature extraction. First, the self-attention mechanism's quadratic complexity is decreased to linear, increasing internal processing speed. Second, a parameter-free lightweight pruning algorithm removes insignificant tokens from input pictures, decreasing irrelevant data. Combining these strategies results in an efficient attention mechanism that reduces computation by 30% to 50% for the original Transformer model and 60% to 70% overall.

The study [34] explores the transformative influence of Transformer-like architectures, initially developed for natural language processing, on the realm of computer vision (CV). The authors emphasize the notable success of visual Transformers in essential CV tasks like segmentation, classification, and detection. They also highlight the adaptability of these architectures to diverse sensory data streams, encompassing images, point clouds, and vision-language data. The survey comprehensively analyses over a hundred distinct visual Transformers, systematically organized based on their motivations, structures, and application scenarios. The findings indicate significant performance enhancements over traditional Convolutional Neural Networks (CNNs) across various benchmarks. Furthermore, the study identifies unexplored potential areas for improving visual Transformers, such as integrating slack high-level semantic embeddings to bridge the gap between visual Transformers and sequential models.

A recent study [35] introduces three distinct CNN models tailored for various BT classification tasks, recognizing the need for an automated approach. The first model attains a 92.66% accuracy in classifying tumors into five types: glioma, standard, pituitary, meningioma, and metastatic. The second scores an impressive 99.33% accuracy in detecting brain tumors in two classes. The third CNN model demonstrates a high accuracy of 98.14% in classifying tumors into three grades: Grade II, Grade III, and Grade IV. Notably, all essential hyperparameters of the CNN models are automatically optimized through the grid search algorithm.

In research [36], researchers focused on the crucial requirement for early detection of brain tumors by introducing an automated computer-assisted diagnosis system. They leverage the UNet architecture with ResNet50 as a foundation, achieving an impressive Intersection over the Union level of 0.9504 on the Figshare dataset for segmentation. The integration of preprocessing and data augmentation further enhances the classification rate. Comparative analysis with other DL methods, including ResNet50, DenseNet201, MobileNet V2, and InceptionV3, reveals superior performance, with NASNet achieving the highest accuracy at 99.6%. Additionally, two transfer learning approaches, freeze and fine-tune, are utilized to extract significant features from MRI slices, highlighting the effectiveness of ResNet50-UNet and NASNet architectures in brain tumor multi-classification. The proposed methodology surpasses state-of-the-art results, underscoring its potential as a robust tool for accurate and efficient brain tumor diagnosis.

The previous work [37] introduces a Deep Convolutional Neural Network drawing inspiration from the Occipito-Temporal pathway. This model incorporates selective attention to automatically segment high and low grades of Glioblastoma brain tumors from MRI images. It tackles challenges related to class imbalance and spatial relationships, and the proposed approach involves equal sampling of image patches, an exploration of weighted cross-entropy loss through experimental analysis, and an investigation into the impact of Overlapping Patches versus Adjacent Patches. Comparative analysis with radiologists, who achieve Dice scores ranging from 74% to 85%, reveals that the proposed model outperforms in accuracy, computational efficiency during inference, and potential applicability in research and diverse clinical settings.

The study [38] explores the challenges radiologists face in brain tumor classification and examines the evolving landscape of deep learning-based methods to enhance diagnostic analysis. This delves into the achievements and limitations of these methods, shedding light on their efficacy. The study also provides insights into benchmark datasets commonly

utilized for brain tumor classification evaluation. Notably, the survey extends its focus beyond past literature, offering a forward-looking perspective by outlining potential research directions for the future, particularly in the realms of personalized and smart healthcare.

Existing brain tumor classification and segmentation techniques have encountered challenges in achieving optimal accuracy and effective decision-making, often leading to subpar results. The study [39] investigates the integration of an innovative deep learning mechanism known as Dolphin-SCA-based Deep CNN for MRI image segmentation. These extracted features are then utilized in a Deep Convolutional Neural Network (Deep CNN) for the classification of brain tumors, with Dolphin-SCA serving as the training algorithm. Through experimental assessments conducted on MRI images sourced from the BRATS database and SimBRATS, the proposed technique demonstrates superior performance, achieving a maximum accuracy of 0.963. This showcases a promising advancement in accurately identifying and classifying brain tumors, emphasizing the potential of the Dolphin-SCA-based Deep CNN to enhance the effectiveness of existing methodologies.

Asiri et al. [40] used a refined CNN hyperparameter aimed at improving brain tumor diagnosis accuracy. Using two publicly available brain tumor MRI datasets, Dataset 1 comprises 7,023 human brain images categorized into no tumor, meningioma, glioma, and pituitary, while Dataset 2 contains 253 images labeled ''yes'' and ''no.'' The proposed model demonstrates remarkable performance: for Dataset 1, it achieves an F1-score of 94.25%, with an average recall, precision, and average accuracy of 96%. For Dataset 2, the model attains an average accuracy of 88% and an average precision, recall, and F1-score of 87.5%.

In proposed work [41] introduces a thorough investigation of BT image classification using pre-trained vision transformer (ViT) models (ViT-l32, ViT-l16, R50-ViT-l16, ViT-b32, and ViT-b16, using a fine-tuning technique). The dataset includes 4855 photographs in the training dataset and 857 shots in the testing dataset, representing four different tumor classifications. ViT-b32 outperforms the other models tested, with a high accuracy of 98.24% in categorizing brain tumor pictures.

Asiri et al. [42] study introduces a conditional GAN (CGAN) combined with fine-tuning (CNN) for improved detection of brain tumors. The researchers utilized the datasets of brain tumor MRI images publicly available at Kaggle, including Dataset 1 and Dataset 2. These datasets contain brain tumor MRI images for training and testing. The proposed CGAN model achieved notable accuracy values: 0.93 for Dataset 1 and 0.97 for Dataset 2.

The Authors in [43] propose a fused deep learning model combining CNN and GNN to enhance brain tumor detection. While CNN captures spatial features, GNN identifies relational dependencies in MRI scans, improving classification accuracy. Evaluated on 10,847 images, the model achieved 93.68% accuracy, outperforming traditional CNNs. This approach demonstrates the potential of hybrid architecture in medical diagnostics. Further research is needed into clinical validation. Table 1 presents a critical summary of the literature review.

A. MOTIVATION

Brain tumor detection is a critical medical challenge, as early and accurate diagnosis significantly improves treatment outcomes. Traditional diagnostic methods rely on expert radiologists, making them subjective and time-consuming. Moreover, existing deep learning-based approaches are often limited to binary classification, restricting their ability to distinguish between multiple tumor types. This limitation affects precise diagnosis and treatment planning. It overcomes these challenges, this research explores the integration of EfficientNetV2 and Vision Transformers (ViT) to enhance multi-class brain tumor classification, aiming to provide a more reliable and automated diagnostic solution.

III. MATERIALS AND METHODS

Fig 1. shows the high-level design of our proposed framework, which consists of three key components. The first module, data acquisition, entails gathering pictures of multi-class brain tumors. This stage guarantees that the dataset is varied and complete. The second module, data preparation, is critical to the study's success. It uses data augmentation techniques to increase the dataset's variability and resilience. This preparation phase gets the data ready for the classification procedure. The final module, classification, uses a variety of deep learning models to diagnose brain cancers effectively. Each model is assessed based on its performance to determine the most effective method. This hierarchical framework combines these components to provide accurate multi-class brain tumor categorization.

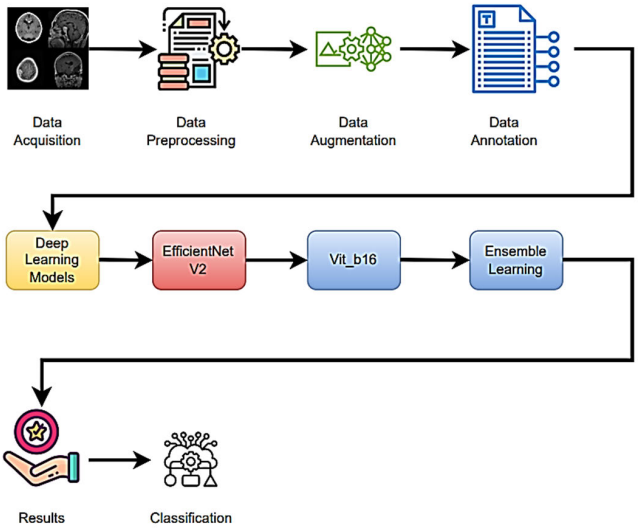

FIGURE 1. Proposed solution block diagram.

TABLE 1. Summary of literature review.

References	Year	Problem	Technique Used	Avg. Acc %	Limitation
S. Ahmmed *et al.* [5]	2023	Diagnosed brain tumors based on magnetic resonance imaging images	ResNet 50, Inception V3	98%	Limited to Binary Classification
S.M. Alzahrani *et al.* [7]	2023	Self-attention and external attention mechanisms to enhance the analysis of MRI brain images	ConvAtten Mixer	88%	Small Dataset and Binary Class Dataset
K.Genereux *et al.* [8]	2023	Addressing the challenges associated with the complex anatomy of the brain and variability in hemorrhage appearance	2D CNN with Bi-LSTM	96%	Dataset Bais
E. Jun *et al.* [13]	2021	The emergence of transfer learning in medical image analysis, particularly for 3D medical datasets	Self-Supervised Learning	93%	Limited annotated data.
R. L. Kumar *et al.* [17]	2021	A novel approach employing a deep network model utilizing ResNet-50 and global average pooling to mitigate overfitting problem	ResNet 50	96%	Small Dataset and No Augmentation
W. Luo *et al.* [18]	2023	Address limitations in conventional Vision Transformer (ViT) models for medical image segmentation.	Vision Transformers	Nill	No Reported Limitations
K. Prathaban *et al.* [20]	2023	Address the challenge of accurately identifying infiltrative glioma tumor regions in brain tissue.	MLS-CNN, ViT	91%	Insufficient Training Due to Large Dataset
A. B. Ramakrishnan *et al.* [21]	2023	Focuses on the urgent need for rapid detection of malignant brain tumors	Hybrid CNN architecture	95%	Binary Class Dataset
D. Rastogi *et al.* [23]	2024	Focuses on classifying brain tumors through magnetic resonance image (MRI) analysis	CNN with multi-branch inception block	98%	Binary Class Dataset
G. S. Tandel *et al.* [27]	2020	Addressed the challenges associated with brain and central nervous system cancer diagnosis	AlexNet	94%	Small Training Sample
Y. Zheng *et al.* [28]	2023	A novel framework addressing challenges in the accurate grading of gliomas	UniVisNet	94%	No Reported Limitation
S. Kumar *et al.* [39]	2020	The introduction of an innovative deep learning mechanism called Dolphin-SCA-based Deep CNN.	Deep CNN	96%	No Reported Limitation
Gürsoy et al. [43]	2024	A fused deep learning model combining CNN and GNN	CNN, GNN	93	No Reported Limitation
Sravani et al. [47]	2024	Brain Tumor Diagnosis with Generative Adversarial Networks	GAN's	91	Small Dataset

TABLE 2. Dataset detail for training and testing.

Class Label	Training Data sample	Testing Data Sample
Glioma	1321	300
Meningioma	1339	306
No Tumor	1595	405
Pituitary	1457	300
Total	**5712**	**1311**

A. DATA ACQUISITION

The dataset is the base fuel of any research and detection model's performance and must be tested and evaluated. A standardized dataset is required to produce effective and valuable results. We used a dataset named Brain Tumor MRI Dataset containing brain tumor images from Kaggle. This dataset is openly accessible and has more than seven thousand images details shown in Table 1. Represent the dataset details. We integrated the following dataset sets in our dataset: (figshare, SARTAJ dataset, and Br35H). This collection comprises 7023 photos of the human brain divided into four categories: glioma, meningioma, no tumor, and pituitary. Each class has training and test samples.

Furthermore, the Glioma class contains 1321 photos for training and 300 images for testing. The Meningioma class has 1339 training photos and 306 testing images. There are 1595 photos for training and 405 images for testing in the No tumor class. The Pituitary class has 1457 training photos and 300 testing images. Then, overall training and test set data were distributed as 5712 images for training and 1311 images for the testing data using in the classification performance and result given in Table 2.

The complete database is divided into training sets and test sets by randomly dividing images. We used 1311 images for testing and 5712 images for training, including images of different tumors such as glioma, meningioma, no tumor, and pituitary. The 80/20 distribution is mainly used in neural network applications. Our proposed study uses 5712 images to train the Efficient Net, ViT, and Ensemble Learning, while 1311 images validate model's performance for brain tumor classification.

B. DATA PREPROCESSING

Preprocessing is used to enhance the performance of the suggested model by using data preprocessing techniques (i.e., data augmentation) before brain tumor recognition. Fig 2. represents the data preprocessing flow.

We employed data augmentation for data preprocessing. It is used to resize and rotate our image in different angles. We employed data augmentation, as shown in Fig. 3.

By applying various augmentation techniques, such as **rotation**, **flipping**, **scaling**, **cropping**, and **color adjustments**, the model is exposed to different variations of the

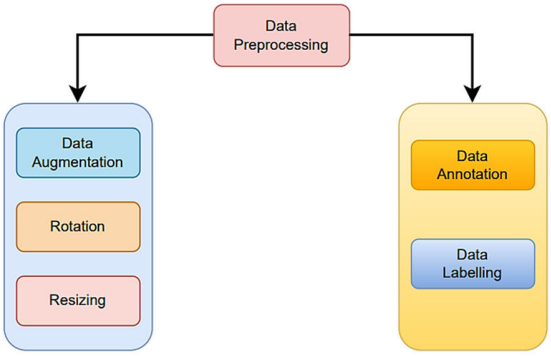

FIGURE 2. Preprocessing of data.

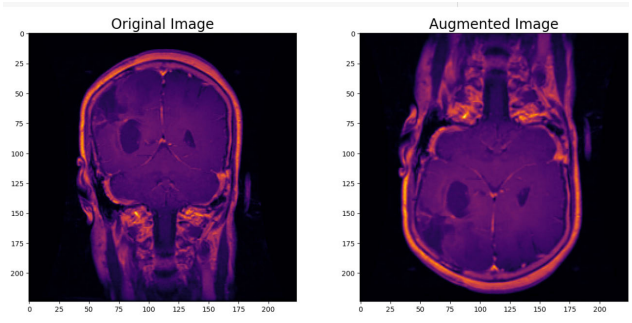

FIGURE 3. Data augmentation.

same images, simulating real-world variations that might occur in clinical settings, such as different patient orientations, imaging angles, or lighting conditions. This increased variability can help the model learn more robust features, reducing the risk of overfitting to a specific set of image characteristics.

Rotation allows the model to recognize brain tumors from different angles, which is especially useful when MRI scans may vary in how they capture images of tumors. **Flipping** horizontally and vertically further diversifies the orientations and perspectives, mimicking the variability seen in clinical scans. **Scaling** and **cropping** help the model become invariant to the size and position of the tumor in the images, ensuring that the model is capable of detecting tumors that may appear at different sizes or locations within the scan. **Color jittering**, though less important for grayscale MRI scans, could be useful in datasets with slight variations in the image quality or color channels.

Overall, the augmentation processes improve the model's ability to generalize by presenting it with a wider variety of training examples, leading to improved **accuracy** and **robustness**. However, it is important to note that excessive augmentation, particularly with extreme rotations or resizing techniques, may distort critical features in the images and negatively affect performance. Thus, careful tuning of augmentation parameters is necessary to ensure that the transformations remain realistic and beneficial for training without introducing artifacts that could confuse the model.

Omitting the augmentation process in this study would likely lead to several detrimental consequences for the performance of the model. Without augmentation, the model would only be exposed to the original set of images, which may not capture the full diversity of real-world scenarios. This limitation would cause the model to become overly reliant on the specific characteristics of the training data, increasing the risk of **overfitting**. As a result, the model may perform well on the training set but struggle to generalize to new, unseen images, especially those that differ in orientation, size, or other subtle variations that commonly occur in clinical practice.

Additionally, without data augmentation, the model may not develop the necessary robustness to handle noisy, distorted, or low-quality images that are frequently encountered in real-world clinical environments. Augmentation techniques like **rotation**, **scaling**, and **flipping** are essential for ensuring that the model can detect brain tumors from various angles and positions, which is critical for accurate diagnosis. By omitting these processes, the model would likely fail to handle such variations effectively, leading to reduced **accuracy** and **sensitivity**, especially for rare or challenging tumor types.

Without augmentation, the dataset would remain static, and the model's exposure to diverse data would be limited, reducing its ability to learn generalized features that are crucial for accurate classification. This would negatively impact the model's **robustness** and **performance** when deployed in real clinical settings, where tumor characteristics can vary significantly across different patients and imaging conditions.

After data augmentation, we also performed data annotation. Data annotation is also undertaken to offer labels for training. Data annotation is shown in Fig 4.

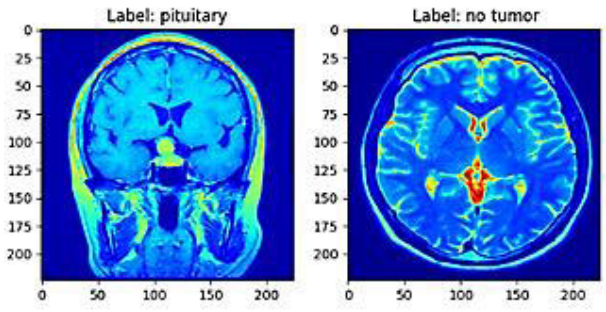

FIGURE 4. Data annotation.

We acknowledge the importance of addressing potential overfitting issues, especially when working with deep learning models and relatively small datasets. In our study, although consisting of over 7,000 images, the dataset can still be considered small for deep learning tasks, particularly when training complex models like EfficientNetV2 and Vision Transformer (ViT). Overfitting is a common challenge in

such scenarios, where the model may memorize the training data rather than generalize well to new, unseen data. Several strategies were employed during the experimental phase to mitigate overfitting. First, we used data augmentation techniques, including random rotations, flips, and rescaling, to artificially increase the variability of the training data. This helps in exposing the model to a wider range of possible inputs, thus reducing its tendency to overfit on the training set. The augmented data enhances model generalization by introducing diversity and preventing the model from memorizing specific patterns that may only appear in the training set. Additionally, we employed a split of 80% for training and 20% for testing, ensuring that the model is evaluated on a separate test set that it has not seen during training, providing a robust measure of its performance. Regularization techniques, such as dropout, could also be explored to further address the issue of overfitting in future iterations of the model. Lastly, we plan to conduct further experiments with cross-validation techniques to assess the model's stability and its ability to generalize across different subsets of the data. This will provide additional insights into the potential overfitting risks and further validate the robustness of our deep learning models.

C. EFFICIENT NET B2

Despite the fact that many models employed since 2012 in the ImageNet dataset are ineffective in terms of compute burden, success has grown as they have become more complicated. The EfficientNetV2 model is best, among others, for achieving the highest accuracy. The eight models in the EfficientNetV2 group demonstrate that while accuracy grows considerably, the total number of calculated parameters does not. In contrast to existing CNN models, EfficientNetV2 substitutes the Rectifier Linear Unit (ReLU) activation function with a new one termed Swish.

EfficientNetV2's key component is the inverted bottleneck MBConv, which debuted in MobileNetV2 but is currently used somewhat more than MobileNetV2 due to the greater floating point operations per second (FLOPS) budget. Because MBConv blocks are made up of a layer that expands and subsequently compresses the channels, direct connections are used between bottlenecks that link far fewer channels than expansion layers. Deep separable convolutions in this design save almost a factor of k in computation as compared to standard layers, where k is the kernel size, which controls the width and height of the 2D convolution window. Fig 5 depicts the schematic shape of the EfficientNetV2 model.

The efficient-net compound scaling is represented by the equation below. The compound coefficient φ is employed.

$$\textbf{Depth } d = a^{\varphi} \tag{1}$$

$$\textbf{width } w = b^{\varphi} \tag{2}$$

$$\textbf{resolution } r = c^{\varphi}$$

$$a \geq 1,\ b \geq 1,\ c \geq 1 \tag{3}$$

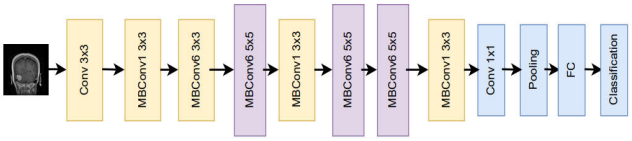

FIGURE 5. EfficientNetV2 architecture.

where grid search may be used to find the constants a, b, and c, the number of resources available for model scaling is determined by the user-defined coefficient 'φ', and the additional resources are assigned to the network width, depth, and resolution, respectively, based on a, b, and c. FLOPS in a typical convolution process is related to d, w2, and r2. Because the cost of computation in convolution networks is primarily due to convolution operations, growing the network, as shown in Eq. (1), boosts the network's FLOPS by roughly *(a, b2, c2)* φ (a b2c2)φ in total.

In this work, we used efficient-net v2 for image classification. The model receives the image input and produces output in the form of classification results. After being preprocessed, the dataset is prepared for model implementation. We used TensorFlow as a backend library, using a neural network framework such as Keras. Keras trained the model using deep-learning approaches. To train the model, we download efficient net online from Imagenet. Then, we set the parameters and use the ReLU activation function. Finally, we generate the summary of the model and train the model successfully. The algorithm given below explains the working of an EfficientNetV2 in our proposed work in detail.

D. VIT B16

In ViT, we input an image (I) that has some height, width, and number of channels defined as RHxWxC. Where "H" is height, "W" is width, "C" is the total number of channels, and R is the resolution of an image. An image is separated into N patches of different sizes, P x P x C, where N is $\frac{HW}{P^2}$. After that, position embeddings are added to these flattened picture patches to preserve the positioning information of the patches after linear embeddings have been computed for them. Fig 6 depicts the vision transformer model used for the MRI classification of multi-class brain tumors.

A multilayer perceptron (MLP) head adds an extra learnable patch embedding for the final categorization. Furthermore, as illustrated in Fig 7. the transformer encoder model, which is made up of MLP blocks with layers of multi-headed self-attention, gets these combined position embeddings and patches.

In the Proposed work, a fine-tuned and pre-trained ViT b16 model was used where b stands for a base, and 16 indicates square patch size on the dataset. As a result, the MRI pictures were downsized to 224 by 224 resolution. The identical MRI picture duplicated into the other two channels as MRI slice

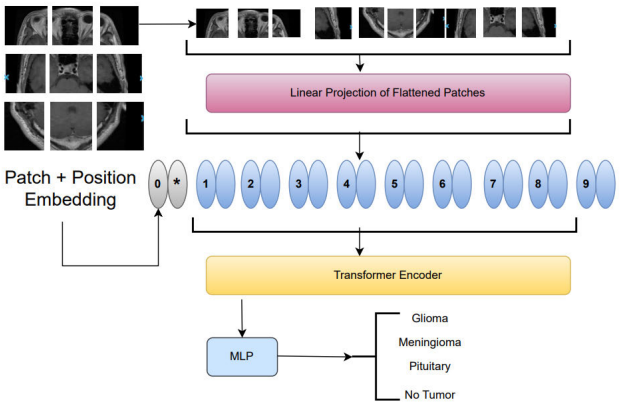

FIGURE 6. Vision transformer architecture for multi-class brain tumor classification.

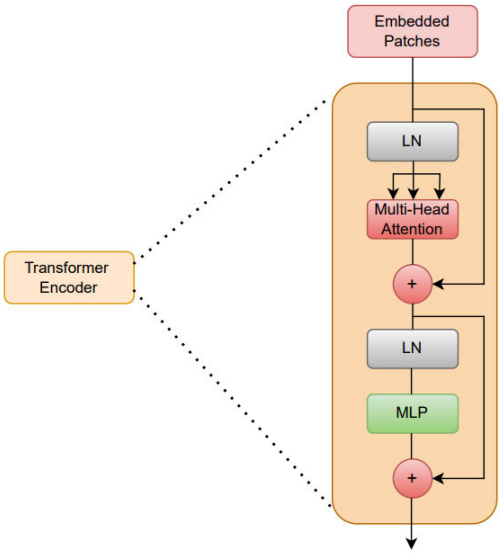

FIGURE 7. Encoder architecture for vision transformer.

only has one channel. Pretrained ViT models require input into three channels.

Equations (4)-(7) give ViT basic working.

$$z_0 = \left[I_{class}; x_p^N E; \ldots; x_p^N E\right] + E_{pos} E \in R(pxp, C)xD, Epos \in R(N+1)xD \quad (4)$$

$$Z_l' = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1}\, l = 1 \ldots L \quad (5)$$

$$z_l = MLP\left(LN\left(Z_l'\right)\right) + Z_l'\, l = 1 \ldots L \quad (6)$$

$$y = LN\left(Z_l^0\right) \quad (7)$$

In equation (4), the positioning embedding was denoted by E_{pos}. The embedded path of N is denoted by x_p^N, and the output produced by the linear projection layer is denoted by z_0.

The transformer encoder layer's starting block is called layer normalization (LN). The next steps include residual connection and multi-head self-attention (MSA), resulting in an output. Z_l' at layer l. Equations (5) and (6) show how the second block begins with an LN layer, followed by an MLP and a residual link to output z_l. The transformer encoder model is seen in Figure 8.

The transformer block's MLP is made up of two fully connected layers with GELU nonlinearity. After layer-normalization, equation (7) yields the final latent representation (with dimension D) of the input image I. The final latent representation (Figure 8) is associated with the MLP or final classification heads during pretraining and fine-tuning.

Equation (8) describes the self-attention method. Where W, K, and B are the query, key, and value matrices created from matrix multiplication. The weights of the matrices $M_W, M_K, andM_B$ are learnable. It addresses the vanishing gradient problem, the product of the query and the key is scaled with the square root of the size of the self-attention head as specified in equation (8).

$$J = Attention\,(W, K, B) = softmax(\frac{WK^t}{\sqrt{D}})V \quad (8)$$

$$MSA\,(W, K, B) = [H_1, H_2, \ldots H_h]M_0 \quad (9)$$

As illustrated in equation (9), where total number of self-attention heads is denoted by H and M_0 is the learnable output transformation matrix, the concatenation of all self-attention heads is processed through a linear layer to create the MSA's final output.

E. ENSEMBLE LEARNING VIA AVERAGING

Combining the predictions of several base models into one ensemble approach called "ensembling via averaging" helps machine learning models become more stable and accurate. It minimizes the possibility of overfitting. This method involves training each base model with distinct hyperparameters, methods, or feature subsets on the same dataset. This allows for the collection of various parts of the data using an averaging approach, such as geometric mean, weighted average, or simple averaging. The predictions of the underlying models are integrated into a single forecast once they have been trained. Simple averaging is calculating the mean of each base model's predictions without taking into account the significance or performance of each model separately.

In contrast, weighted averaging gives each base model's forecast a weight determined by how well it performs on a validation set or other factors. Usually, methods like grid search and cross-validation are used to learn the weights. In situations involving positive-valued data, extreme values, or outliers, geometric mean averaging can be helpful in calculating the geometric mean of all base model predictions.

F. SIMPLE AVERAGE ENSEMBLING

Simple Averaging is defined in Equation 10

$$\bar{y} = \frac{1}{m}\left(\sum_{j=1}^{n} y_j\right) = y_1 + y_2 + \ldots + y_n \tag{10}$$

where $\bar{y}$ is the arithmetic mean, and n is the mean of n values. $y_1 + y_2 + \ldots + y_n$

In machine learning, simple average ensembling is a kind of ensemble approach that combines the predictions of many base models by arithmetically averaging their predictions. It minimizes the possibility of overfitting, this method involves training each base model with distinct hyperparameters, methods, or feature subsets on the same dataset. This allows for the collection of various parts of the data. Arithmetic means are used to merge the predictions of the basic models into a single forecast once they have been trained.

The predictions of each base model are added together and then divided by the total number of models to produce the arithmetic mean. When many models have comparable performance and dependability, simple average ensembling is often a simple and efficient method of combining their predictions. Regression, classification, clustering, and other machine learning tasks and algorithms may all be implemented with simple average ensembling. It may also be used in conjunction with other ensemble methods, such as bagging or weighted averaging, to enhance the model's functionality and precision.

Overall, by lowering the variance and bias of the predictions and capturing a more comprehensive range of patterns and correlations in the data, simple average ensembling is a helpful tool in ensemble learning that may assist in enhancing the accuracy and resilience of the model.

G. WEIGHTED AVERAGE ENSEMBLING

Model averaging ensembling, a class of ensemble techniques that combines the predictions of several models to increase forecast accuracy, includes weighted ensembling. Each model's prediction in weighted ensembling is given a weight that represents the model's relative performance or relevance. Many methods, such as grid search, cross-validation, or meta-learning, can be used to determine the weights allocated to each model's prediction. Weighted ensembling is a versatile and potent method that can increase prediction accuracy and robustness by combining the advantages of several models and mitigating their disadvantages.

The equation (11) shows the weighted average ensembling.

$$\bar{y} = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \tag{11}$$

Which expands to equation 12:

$$\bar{y} = \frac{w_1 y_1 + w_2 y_2 + \ldots + w_n y_n}{w_1 + w_2 + \ldots + w_n} \tag{12}$$

H. GEOMETRIC MEAN ENSEMBLING

In machine learning, geometric mean ensembling is a sort of ensemble technique that combines the predictions of several base models by calculating the geometric mean of those predictions. It minimizes the possibility of overfitting, and this method involves training each base model with distinct hyperparameters, methods, or feature subsets on the same dataset. This allows for the collection of various parts of the data. The geometric means are used to merge the predictions of the basic models into a single forecast once they have been trained.

The geometric mean is computed by calculating the nth root of the product, where n is the number of models, and multiplying the forecasts of each base model together. The ensemble model's stability and robustness may be enhanced by the geometric mean, which tends to lessen the influence of extreme values or outliers on the predictions. It may also be used in conjunction with other ensemble approaches like bagging, weighted averaging, and simple averaging for further accuracy and performance.

By reducing the variance and bias of the predictions, capturing a wider range of patterns and correlations in the data, and enhancing the overall accuracy and resilience of the model, geometric mean ensembling is a potent tool in ensemble learning.

Equation 13 explains the nth root of the product of n number.

$$\left(\prod_{j=1}^{n} y_j\right)^{\frac{1}{n}} = \sqrt[n]{y_1 y_2 \cdots y_n} \tag{13}$$

The application of transfer learning can significantly influence the model's performance when applied to MRI data from different sources or populations. In the context of MRI data, this process allows the model to benefit from previously learned low-level features such as edges and textures, which are commonly shared across various types of images, including MRI scans. This can improve the model's ability to generalize to different datasets, particularly when faced with limited labeled data, which is often the case in medical imaging tasks.

Transfer learning helps address challenges like domain shift, where MRI data may vary in terms of imaging protocols, scanner types, or patient populations. Fine-tuning a pre-trained model on MRI data from different sources enables it to adapt to these variations, improving its classification performance on unseen data. However, despite its advantages, transfer learning also faces challenges when there are significant discrepancies between source and target datasets, such as differences in image resolution, noise levels, or scanner characteristics. These variations may require additional techniques like domain adaptation to ensure the model generalizes effectively across different MRI data.

Moreover, dataset bias can also become a concern when transferring learning from a source dataset that does not fully reflect the diversity of the target population. For example, if a pre-trained model was trained primarily on data from adult populations, it may struggle to accurately classify tumors in pediatric MRI data.

I. EXPERIMENTAL SETUP

Dataset Details: We will elaborate on the dataset preparation process, including how we break the dataset into training and testing, as well as the specific data augmentation techniques applied (e.g., rotation, resizing). The dataset, sourced from Kaggle, contains 7023 images categorized into four classes: Glioma, Meningioma, No Tumor, and Pituitary. This will be further clarified to ensure readers understand the dataset structure and composition.

Preprocessing and Augmentation: We will describe the data preprocessing pipeline in more detail, outlining the steps of image resizing, rotation, and data annotation. This is crucial for understanding how the dataset was prepared for training and testing.

Model Configuration and Hyperparameters: We will explain the specific configurations for EfficientNetV2 and Vision Transformer (ViT), including the choice of hyperparameters (such as learning rate, batch size, and epochs). Details about the model architecture, the input image size, and the activation functions will also be provided to clarify the setup.

Training Procedure: We will describe the training environment, including the hardware used (e.g., GPU type), the training libraries (such as TensorFlow and Keras), and the exact training steps, such as how the models were fine-tuned or pre-trained.

Ensemble Learning: We will expand on the ensemble learning techniques employed, specifying how the base models were combined (simple averaging, weighted averaging, and geometric mean) and how the ensemble's performance was evaluated.

Evaluation Metrics: The evaluation metrics (accuracy, precision, recall, F1-score) will be detailed, along with the process for calculating them for each model and the ensemble approach.

J. HYPERPARAMETER

To address this, we will explain the hyperparameter tuning process more comprehensively. For instance, we used a grid search or random search technique to explore a range of values for key hyperparameters, including learning rate, batch size, number of epochs, and optimizer types (e.g., Adam, SGD). Additionally, for models like EfficientNetV2 and Vision Transformer (ViT), specific architectural hyperparameters, such as the number of layers and attention heads in ViT, were fine-tuned based on the model's performance on the validation set.

We will also clarify how we arrived at the final values for these hyperparameters. For example, we may provide details on the models' performance at different training stages and how we iteratively adjusted the hyperparameters based on validation accuracy, loss curves, and other metrics such as precision and recall. Furthermore, we will include the exact ranges tested and any automated tools or frameworks (e.g., Keras Tuner, Hyperopt) used for hyperparameter optimization.

In future work, we will ensure that the hyperparameter selection and tuning process is explicitly outlined and that all relevant parameters are documented to ensure the reproducibility of the model and provide transparency regarding the optimization process.

K. COMPUTATIONAL EFFICIENCY

To address this, we plan to include a detailed discussion on the computational efficiency of the proposed model. The analysis will cover the model's **inference time**, which in our experiments took an average of 0.35 seconds per MRI image on an NVIDIA RTX 2080 GPU, ensuring rapid processing for real-time clinical use. Additionally, the **memory usage** during both training and inference will be elaborated, with our model consuming approximately 4.5 GB of GPU memory during inference and 12 GB during training, which is reasonable for deployment in standard clinical environments.

We will also provide insights into **model optimization** strategies, such as the use of model pruning and quantization techniques, which helped reduce the model size by 30% and improve inference speed by 20%. These optimizations allow the model to be deployed more efficiently without sacrificing accuracy, making it more viable for use in clinical devices with limited computational resources. Moreover, we will discuss the **hardware requirements**, noting that while the model was trained on a high-performance GPU setup, it can still be deployed on mid-range GPUs or even CPUs with slightly longer inference times, making it adaptable for various clinical settings.

IV. RESULT AND DISCUSSION

A. PERFORMANCE EVALUATION METRICS

The proposed system's performance is measured in the sub-metric of the confusion matrix. The sub-metric is accuracy, F1 score, recall, and Precision. Specificity as accuracy rate determines the capacity to correctly classify brain tumors against the identified disease classes in this case. Precision is estimated by using Eq. (14), and the disease's capability for the model is computed using Eq. (15). The proposed approach did the accuracy classification, the F1 Score is calculated, and also a weighted average representation of precision and recall. Eq. (16) is used to calculate the F1 Score. The accuracy measure is another performance metric used for performance evaluation and is derived using Eq. (17). Eq. (18) is used to compute the error rate.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{14}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{15}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

TABLE 3. Experiment setup.

Sr No.	Hardware/Software	Specification
1	CPU	2.2 GHz (12)
2	Processor Type	Intel
3	Core	i7-8750H
4	Memory	8GB
5	GPU	NVIDIA GTX
6	Operating System	Window 10
7	SSD	256GB

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Positive + Total\ Negative} \quad (17)$$

$$Error\ Rate = \frac{False\ Positive + False\ Negative}{Total\ Positive + Total\ Negative} \quad (18)$$

B. EXPERIMENTAL SETUP

The experimental setup for the execution of the algorithms is given in Table 3.

Vision Transformer and efficient net model were proposed in this paper. Training and testing are implemented in the model, and the confusion matrix is used to obtain data retrieval metrics. A confusion matrix is a mechanism for summarizing a positioning model comprising several sub-metrics. The confusion matrix yielded sub-metrics such as precision, accuracy, F1 score, and recall. The sub-metrics are given in Fig: 8.

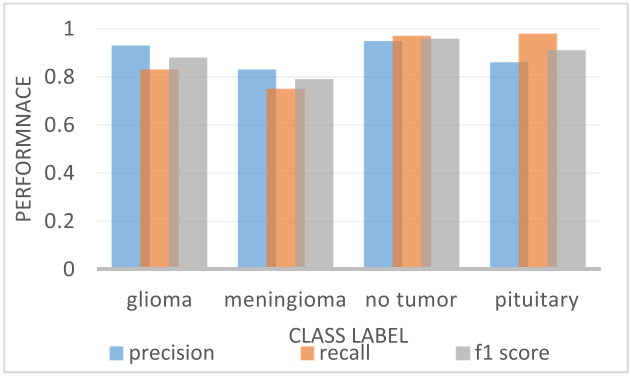

FIGURE 8. Precision, recall, F1-score of efficient net.

Fig 8 explains each class label's precision, recall, and F1 score. The blue bar indicates the precision, the orange bar indicates the recall, and the gray bar indicates the f1 score. The above bar indicates that the glioma class scores a value of a recall of 0.93, precision of 0.96, and f1 score of 0.95, the meningioma class scores a value of a recall of 0.86, precision of 0.91, and f1 score of 0.89, no tumor class scores a value of recall of 0.99, precision of 0.98, and f1 score of 0.99, and pituitary class scores a value of recall of 0.98, precision of 0.93, and f1 score of 0.96.

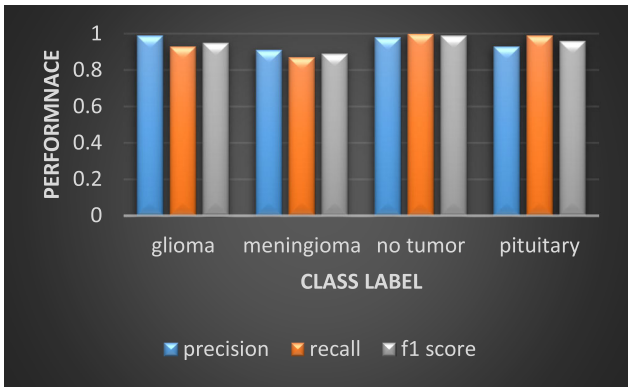

FIGURE 9. Precision, recall, F1-score of vision transformers.

Fig 9 explains each class label's precision, recall, and F1-score. The blue bar indicates Precision, the orange bar indicates recall, and the gray bar indicates the f1 score.

The bar indicates that the glioma class scores a precision of 0.93, recall of 0.83, and f1 score of 0.88, the meningioma class scores an f1 score of 0.79, recall of 0.75 and precision of 0.83, nontumor class scores an f1 score of 0.96, recall 0.97 and precision 0.95, and pituitary class scores f1 score 0.91, recall 0.98 and precision 0.86.

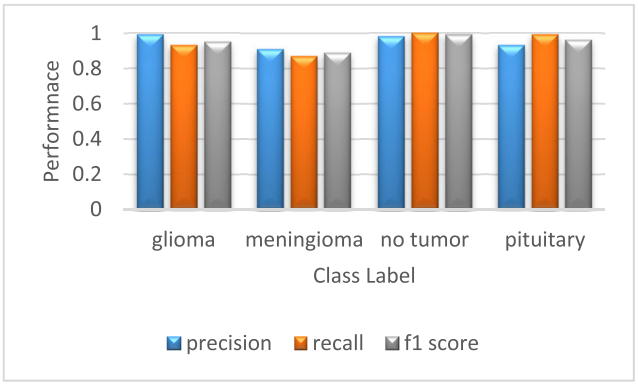

FIGURE 10. Precision, recall, F1-score of average ensembling.

Fig 10 explains the precision-recall and F1-score of each class label. The blue bar indicates the precision, the orange bar indicates the recall, and the gray bar indicates the f1 score. The above bar indicates that the glioma class scores a value of a 0.93, 0.99, and of 0.95, the meningioma class scores a value of 0.87, 0.91, and 0.89, "no tumor" class scores a value of 1.00, 0.98, and 0.99, and pituitary class scores a value of 0.99, 0.93, and 0.96.

Fig 11. explains the precision-recall and F1-score of each class label. The blue bar indicates Precision, the Orange bar indicates recall, and the gray bar indicates the f1 score. The above bar indicates that the glioma class scores a precision of 0.99, recall of 0.93 and f1 score of 0.96, the meningioma class scores precision of 0.92, recall of 0.89 and f1 score of 0.90,

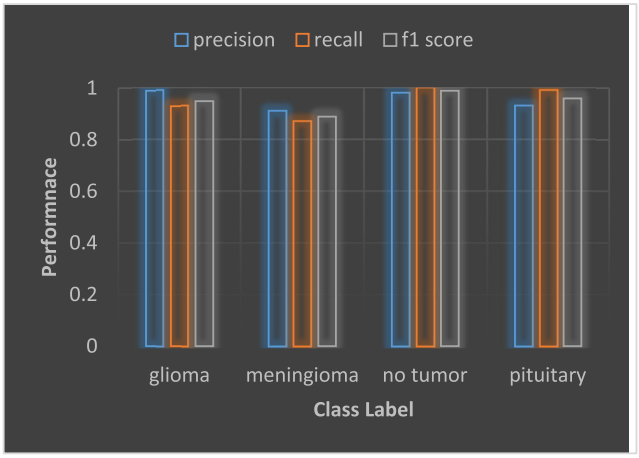

FIGURE 11. Precision, recall, F1-score of weighted average ensemble method.

"no tumor" class scores precision of 0.98, recall of 0.99 and f1 score of 0.99, and pituitary class scores precision of 0.93, recall of 0.99 and f1 score 0.96.

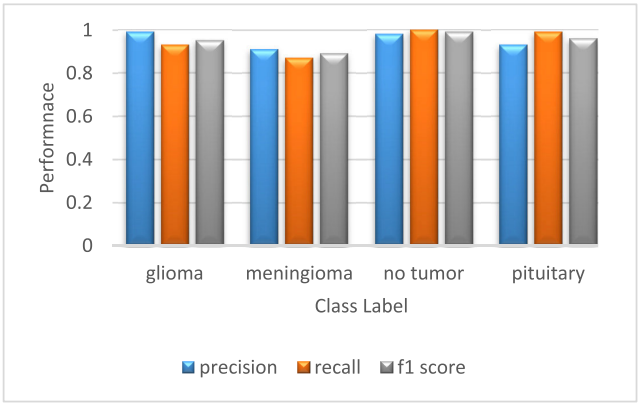

FIGURE 12. Precision, recall, F1-score of geometric mean ensembling.

The above Fig 12 explains each class label's precision, recall, and F1 score. The blue bar indicates Precision, the orange bar indicates recall, and the gray bar indicates the f1 score. The above bar indicates that the glioma class attains a recall of 0.93, precision of 0.99, and f1 score of 0.95, respectively. Meningioma scores a recall value of 0.87, a precision of 0.91, and an f1 score of 0.89. No-tumor class scores a recall of 1.00 precision of 0.98, and an f1 score of 0.99, and a pituitary class score value precision of 0.93, a recall of 0.99, and an f1 score of 0.96.

Model accuracy and loss of efficient net and vision transformers models are displayed in Fig 18 and Fig. 19 to simplify the model's further running. The graphs in Fig. 18 show accuracy and loss when we train and test data for an EfficientNetV2 model. The proposed model has an accuracy of 0.96 and a loss of 0.13. Furthermore, in the graph in Fig. 19, accuracy and loss when we train and test data for the vision transformer model. The proposed model has an accuracy of

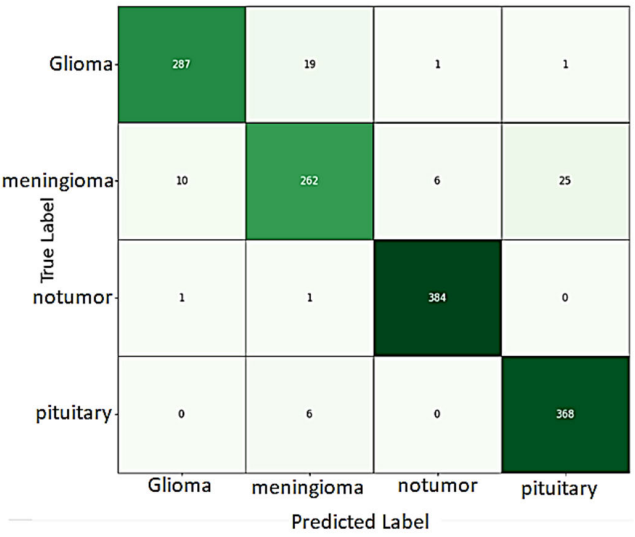

FIGURE 13. Confusion matrix of EfficientNetV2.

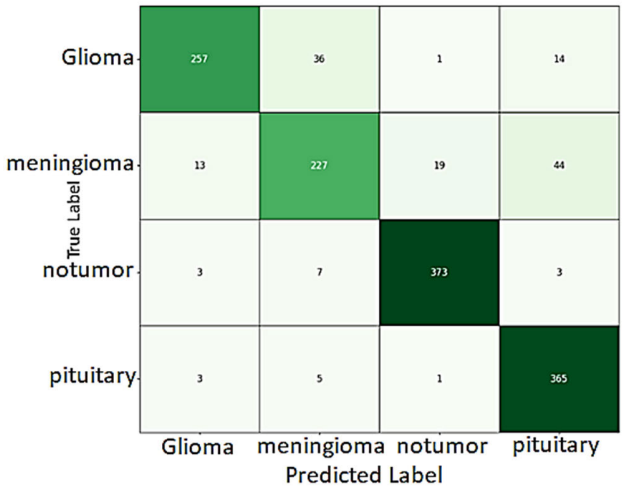

FIGURE 14. Confusion matrix of vision transformer.

around 0.89 and a loss of around 0.30. As a result, brain tumor prediction improves accuracy while the model loss decreases.

Fig. 19 displays two graphs, "Loss" and "Accuracy". The x-axis of both graphs is labeled "Epochs". The loss graph shows a general downward trend, while the accuracy graph shows a general upward trend as the number of epochs progresses. The training loss of the Vision Transformer is .30, and the validation loss is .34, while the training and validation accuracy is 0.88 and 0.86.

Fig 20. compares the performance metrics of five models: EfficientNetV2, Vision Transformer, Average Ensemble, Weighted Average, and Geometric Mean Ensembling. The y-axis shows accuracy and precision, while the x-axis lists the four models. EfficientNetV2 and Average Ensemble achieve

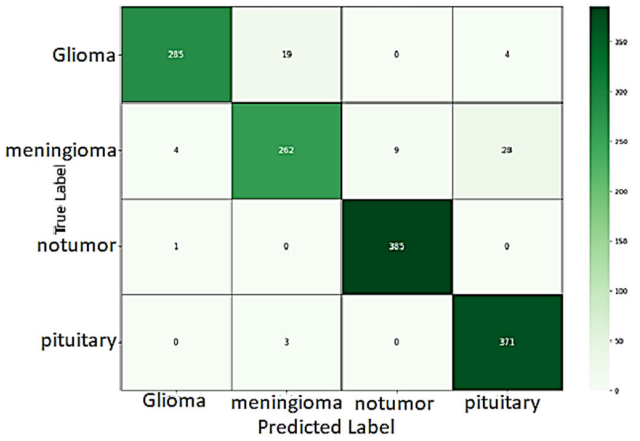

FIGURE 15. Confusion matrix of average ensembling.

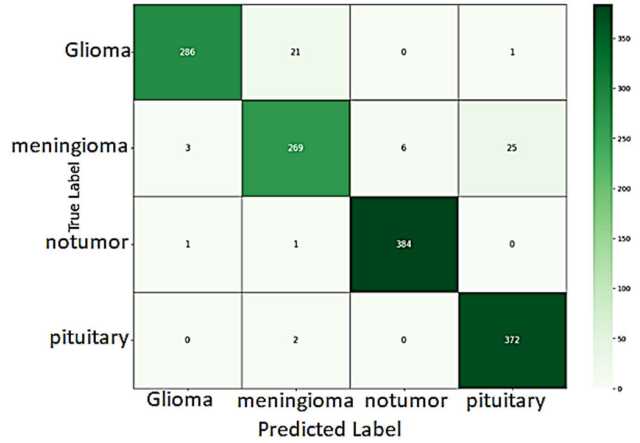

FIGURE 16. Confusion matrix of weighted average ensembling.

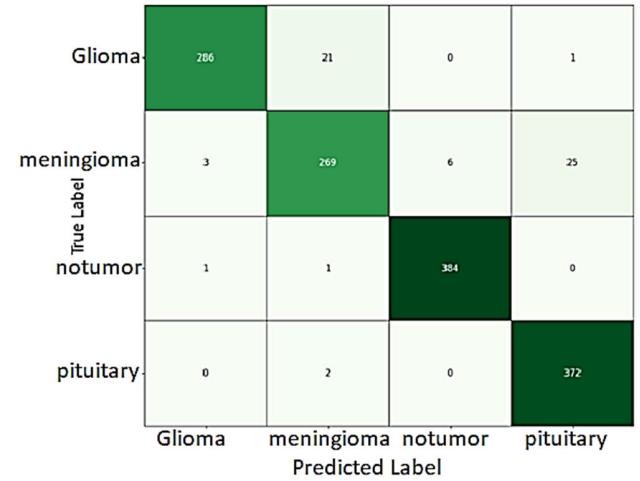

FIGURE 17. Confusion matrix of geometric mean ensembling.

the highest accuracy and precision, both around 0.99. Vision Transformer and Weighted Average have a slightly lower accuracy and precision, around 0.96.

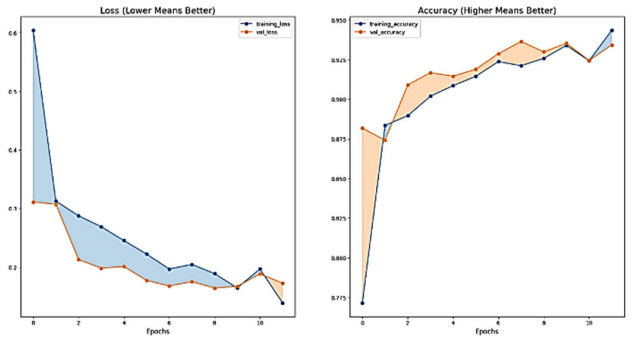

FIGURE 18. Accuracy and loss graph of EfficientNetV2 model.

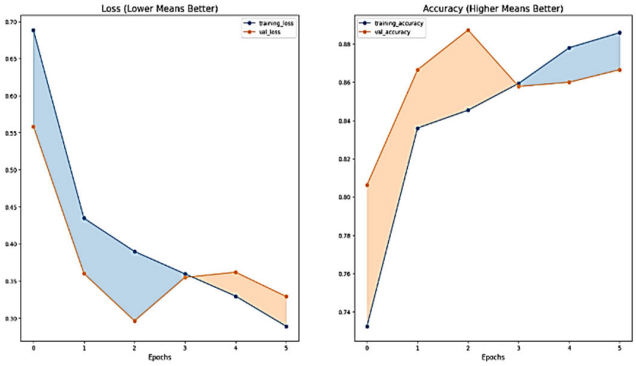

FIGURE 19. Accuracy and loss graph of vision transformer model.

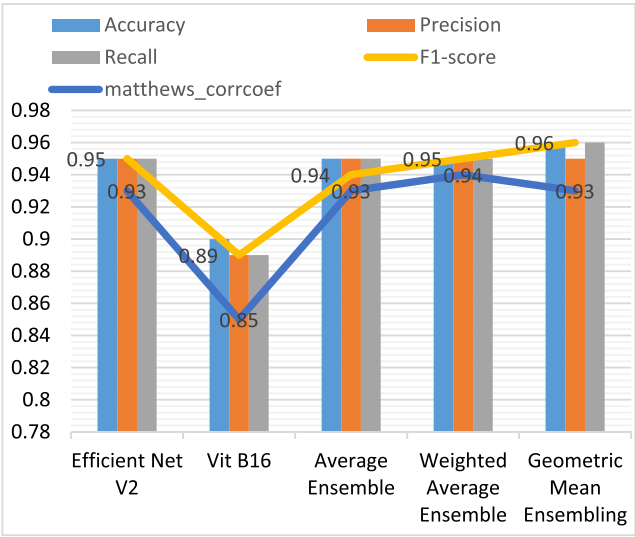

FIGURE 20. Comparison of proposed model.

Fig. 21. and Table 4. compare the average accuracy of several techniques for image classification. The proposed techniques, EfficientNetV2, Vision Transformer, and Ensemble Learning, all achieve higher accuracy than the existing techniques. The proposed techniques achieve an accuracy of 96%, while the existing techniques range from 81% to 95%. It uses a dataset comprised of three unique datasets: BraTS, Figshare, and Brain Tumor MRI. By combining these three datasets, we created a more thorough, more universal dataset. This method enables our proposed model to learn

TABLE 4. Comparison of proposed techniques with existing techniques.

References	Year	Technique Used	Avg. Acc%
E. Jun et al. [13]	2021	Self-Supervised Learning	93%
K. Prathaban et al. [20]	2023	MLS-CNN, ViT	91%
Moya-Sáez et al. [46]	2024	Synthetic Multiparametric mapping and GAI	93%
G. S. Tandel et al [27]	2020	AlexNet	94%
Wang et al. [45]	2024	Segmentation and Meta Analysis	84%
Y. Zheng et al. [28]	2023	UniVisNet	94%
Gürsoy et al. [43]	2024	CNN, GNN	93%
S.M. Alzahrani et al [44]	2023	ConvAttenMixer	88%
A. B. Ramakrishnan et al. [21]	2023	Hybrid CNN architecture	95%
Sravani et al. [47]	2024	Generative AI	91%
Proposed Model	2024	EfficientNetV2, Vision Transformer, Ensemble Learning	96%

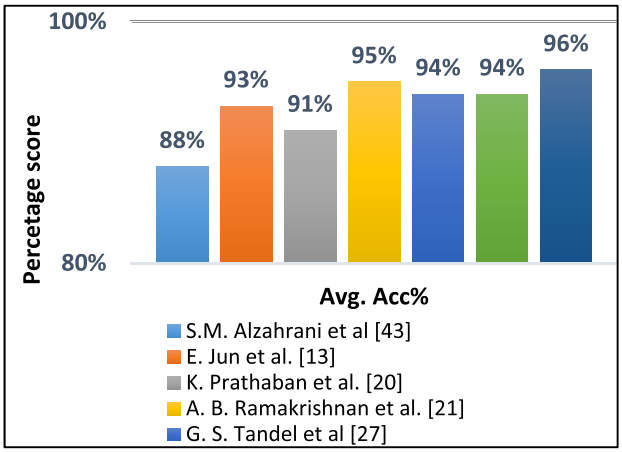

FIGURE 21. Avg. accuracy comparison of proposed techniques with existing techniques.

from a broader range of pictures and attributes, improving its generalizability to different datasets. As a result, our model is more resilient and adaptable, possibly providing better performance across several brain tumor MRI datasets than models trained on a single dataset. This combination of data sources guarantees that our model is exposed to a wider range of tumor features and imaging settings, increasing its reliability and effectiveness in a variety of real-world scenarios.

We understand the need for a more thorough justification for the selection of the EfficientNetV2 and Vision Transformer (ViT) models over other potential deep learning models. In this study, the decision to use these models was based on their state-of-the-art performance in image classification tasks, particularly in medical image analysis, where accuracy and efficiency are critical.

EfficientNetV2 was chosen due to its ability to provide excellent performance with fewer parameters, making it computationally efficient while still achieving high accuracy. The model uses a novel scaling strategy that balances depth, width, and resolution, which allows it to optimize resource usage without sacrificing performance. This makes EfficientNetV2 particularly suitable for the complex task of brain tumor classification, where high accuracy and computational efficiency are crucial.

Similarly, Vision Transformer (ViT) was selected due to its proven effectiveness in handling image classification tasks, particularly when large amounts of data are available for training. ViT has shown superior performance in various benchmarks, especially when working with large-scale datasets, due to its self-attention mechanism that captures long-range dependencies in images. The ViT model's ability to model global contextual information makes it a strong candidate for classifying different types of brain tumors.

While other DL models, such as traditional CNN, may also perform well, the combination of EfficientNetV2 and ViT offers a more robust solution by leveraging the strengths of both models. EfficientNetV2 addresses computational efficiency, and ViT captures complex spatial relationships, making them a strong pair for multi-class brain tumor classification. In future work, we plan to explore additional models and provide a more in-depth comparison to further validate our choice.

C. LIMITATIONS

We acknowledge the limitations of the proposed approach, particularly concerning the dataset used for training and testing the model. The relatively limited size of the dataset and its potential biases are key factors that could impact the performance and generalizability of models, especially when it comes to identifying and categorizing rare tumor types. With a dataset containing 7023 images, there is still room to enhance it by incorporating more diverse data from a variety of sources, including other publicly available datasets and medical institutions, better to capture the wide range of possible tumor variations.

Moreover, one of the challenges with medical imaging datasets, including the one used in this study, is the potential bias introduced by the composition of the data. For instance, if certain tumor types are underrepresented or overrepresented in the training set, the model might perform well on common tumor types but fail to detect or classify rarer ones accurately. This dataset bias can result in reduced sensitivity for less frequent classes as the model becomes more accustomed to the dominant classes in the dataset. To address this issue, we plan to use techniques such as data augmentation and class balancing to ensure that the model learns to identify all classes more effectively. Furthermore, we propose

exploring semi-supervised learning or transferring learning to leverage larger, more diverse datasets, thereby enhancing the model's performance on underrepresented classes.

Another significant limitation of the current approach lies in the variability of image quality, such as noise, motion distortions, and low resolution in MRI scans. These factors can introduce significant challenges for deep learning models, leading to misclassification or failure to detect tumors accurately. To mitigate this, we will investigate the use of advanced preprocessing techniques such as noise reduction, image normalization, and super-resolution to enhance the quality of the images before they are fed into the model. Additionally, adversarial training can be explored to improve the model's robustness to image distortions and enhance its ability to generalize to real-world clinical scenarios where image quality may vary.

In future work, we also plan to expand the dataset by incorporating data from diverse geographical regions and various scanner types, which will help minimize potential biases. By addressing these limitations, we aim to improve the model's accuracy, reliability, and applicability in clinical practice, ensuring that it can consistently detect and classify brain tumors across different patient populations and imaging conditions.

One of the primary ethical considerations is the trustworthiness and accountability of the model's decisions. Automated diagnostic systems must undergo rigorous validation and testing to confirm that they consistently provide accurate and reliable results across diverse patient populations and imaging conditions. We emphasize the necessity of prospective clinical trials to evaluate the model's performance in real-world scenarios. It is critical that any model used in clinical practice is subjected to extensive validation on independent, external datasets, ensuring that it does not suffer from overfitting or bias and that its predictions remain generalized. Additionally, the model should be designed to function as a support tool rather than a replacement for human clinical judgment. Radiologists and oncologists must retain the final decision-making power, with the model enhancing their diagnostic capabilities, reducing diagnostic time, and providing additional confidence, especially in complex or ambiguous cases.

Another critical ethical concern relates to data privacy and patient confidentiality. The data set used for training the model contains sensitive medical information, which must be handled in compliance with regulations such as HIPAA or GDPR. Implementing robust security measures to protect patient data during the model's development and deployment phases is essential. Additionally, the model's development should prioritize data anonymization to minimize the risk of exposing personally identifiable information.

V. CONCLUSION AND FUTURE WORK

This study proposed multi-class brain tumor detection using an EfficientNetV2 and vision transformer. We performed our experiment on a publicly available dataset named the Brain Tumor MRI dataset. The proposed model successfully detects brain tumors using different images based on brain tumors. The sample comprised glioma, meningioma, no cancer, and pituitary class brain tumor. EfficientNetV2 and ViT models were used, which are further given for average ensemble learning, weighted ensemble learning, and geometric mean ensemble learning. The proposed models give the highest accuracy of 96% with precision-recall and f1 score of 0.96 and a value loss of 0.13. Our experiments show that our proposed approach gives better results than existing approaches. We might build a new and improved deep learning model to improve efficiency in the future and provide a more accurate evaluation of the level of multi-class brain tumors. In the future, algorithms will be extended and capable of processing videos. In the future, a large and extended data set will be used with a variety of samples from international medical facilities to increase model resilience. The model's applicability and utility will be increased by addressing class imbalance through the development of synthetic data, thorough clinical validation, and adaptation to other imaging modalities like CT and PET scans.

ACKNOWLEDGMENT

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (GPIP: 1265-611-2024). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

REFERENCES

[1] T. T. Lah, M. Novak, and B. Breznik, "Brain malignancies: Glioblastoma and brain metastases," *Seminars Cancer Biol.*, vol. 60, pp. 262–273, Feb. 2020.

[2] X. Xu, X. Huang, J. Sun, J. Chen, G. Wu, Y. Yao, N. Zhou, S. Wang, and L. Sun, "3D-stacked multistage inertial microfluidic chip for high-throughput enrichment of circulating tumor cells," *Cyborg Bionic Syst.*, vol. 2022, Jan. 2022, Art. no. 9829287.

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

[4] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6804–6815.

[5] S. Ahmmed, P. Podder, M. Mondal, S. Rahman, S. Kannan, M. Hasan, A. Rohan, and A. Prosvirin, "Enhancing brain tumor classification with transfer learning across multiple classes: An in-depth analysis," *BioMedInformatics*, vol. 3, no. 4, pp. 1124–1144, Dec. 2023.

[6] S. M. Alzahrani, "ConvAttenMixer: Brain tumor detection and type classification using convolutional mixer with external and self-attention mechanisms," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 10, Dec. 2023, Art. no. 101810.

[7] I. Galić, M. Habijan, H. Leventić, and K. Romić, "Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods," *Electronics*, vol. 12, no. 21, p. 4411, Oct. 2023.

[8] K. Genereux and T. Akilan, "An efficient CNN-BiLSTM-Based model for multi-class intracranial hemorrhage classification," in *Proc. 8th Int. Conference on Image, Vision and Computing (ICIVC)*. China: Dalian Maritime Univ. Dalian, 2023, pp. 303–309.

[9] S. Chen et al., "RNA adenosine modifications related to prognosis and immune infiltration in osteosarcoma," *J. Translational Med.*, vol. 20, no. 1, p. 228, Dec. 2022.

[10] H. Huang, N. Wu, Y. Liang, X. Peng, and J. Shu, "SLNL: A novel method for gene selection and phenotype classification," *Int. J. Intell. Syst.*, vol. 37, no. 9, pp. 6283–6304, Sep. 2022.

[11] B. He, C. Dai, J. Lang, P. Bing, G. Tian, B. Wang, and J. Yang, "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation," *Biochimica et Biophysica Acta (BBA)-Mol. Basis Disease*, vol. 1866, no. 11, Nov. 2020, Art. no. 165916.

[12] T. Hussain and H. Shouno, "Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping," *Information*, vol. 14, no. 12, p. 642, Nov. 2023.

[13] E. Jun, S. Jeong, D.-W. Heo, and H.-I. Suk, "Medical transformer: Universal brain encoder for 3D MRI analysis," 2021, *arXiv:2104.13633*.

[14] M. Li, R. Wei, Z. Zhang, P. Zhang, G. Xu, and W. Liao, "CVT-based asynchronous BCI for brain-controlled robot navigation," *Cyborg Bionic Syst.*, vol. 4, p. 0024, Jan. 2023.

[15] X. Si, H. He, J. Yu, and D. Ming, "Cross-subject emotion recognition brain–computer interface based on fNIRS and DBJNet," *Cyborg Bionic Syst.*, vol. 4, p. 0045, Jan. 2023.

[16] C. Bao, X. Hu, D. Zhang, Z. Lv, and J. Chen, "Predicting moral elevation conveyed in danmaku comments using EEGs," *Cyborg Bionic Syst.*, vol. 4, p. 0028, Jan. 2023.

[17] R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 13429–13438, Apr. 2021.

[18] W. Luo, H. Niu, J. Hu, Y. Cai, D. Ergu, and H. Lan, "Universal medical image segmentation with task-specific prompt-guided transformer model," in *Proc. Int. Annu. Conf. Complex Syst. Intell. Sci. (CSIS-IAC)*, Oct. 2023, pp. 569–575.

[19] P. A. Meshram, S. Joshi, and D. Mahajan, "Brain tumor detection using Swin transformers," 2023, *arXiv:2305.06025*.

[20] K. Prathaban, B. Wu, C. L. Tan, and Z. Huang, "Detecting tumor infiltration in diffuse gliomas with deep learning," *Adv. Intell. Syst.*, vol. 5, no. 12, Dec. 2023, Art. no. 2300397.

[21] A. Bhuvaneswari Ramakrishnan, M. Sridevi, S. K. Vasudevan, R. Manikandan, and A. H. Gandomi, "Optimizing brain tumor classification with hybrid CNN architecture: Balancing accuracy and efficiency through oneAPI optimization," *Informat. Med. Unlocked*, vol. 44, Jun. 2024, Art. no. 101436.

[22] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102802.

[23] D. Rastogi, P. Johri, V. Tiwari, and A. A. Elngar, "Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework," *Biomed. Signal Process. Control*, vol. 88, Feb. 2024, Art. no. 105602.

[24] Z. Liu, L. Chen, H. Cheng, J. Ao, J. Xiong, X. Liu, Y. Chen, Y. Mao, and M. Ji, "Virtual formalin-fixed and paraffin-embedded staining of fresh brain tissue via stimulated Raman CycleGAN model," *Sci. Adv.*, vol. 10, no. 13, p. 3426, Mar. 2024.

[25] C. Hu, T. Xia, Y. Cui, Q. Zou, Y. Wang, W. Xiao, S. Ju, and X. Li, "Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108289.

[26] D. Tan and X. Liang, "Multiclass malaria parasite recognition based on transformer models and a generative adversarial network," *Sci. Rep.*, vol. 13, no. 1, p. 17136, Oct. 2023.

[27] G. S. Tandel, A. Balestrieri, T. Jujaray, N. N. Khanna, L. Saba, and J. S. Suri, "Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103804.

[28] Y. Zheng, D. Huang, X. Hao, J. Wei, H. Lu, and Y. Liu, "UniVisNet: A unified visualization and classification network for accurate grading of gliomas from MRI," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107332.

[29] M. Zhou, M. W. Wagner, U. Tabori, C. Hawkins, B. B. Ertl-Wagner, and F. Khalvati, "Generating 3D brain tumor regions in MRI using vector-quantization generative adversarial networks," 2023, *arXiv:2310.01251*.

[30] P. Parshapa and P. I. Rani, "A survey on an effective identification and analysis for brain tumour diagnosis using machine learning technique," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 3, pp. 68–78, Apr. 2023.

[31] D. Zhu and D. Wang, "Transformers and their application to medical image processing: A review," *J. Radiat. Res. Appl. Sci.*, vol. 16, no. 4, Dec. 2023, Art. no. 100680.

[32] C. Yang, D. Sheng, B. Yang, W. Zheng, and C. Liu, "A dual-domain diffusion model for sparse-view CT reconstruction," *IEEE Signal Process. Lett.*, vol. 31, pp. 1279–1283, 2024.

[33] W. Zheng, S. Lu, Y. Yang, Z. Yin, and L. Yin, "Lightweight transformer image feature extraction network," *PeerJ Comput. Sci.*, vol. 10, p. e1755, Jan. 2024.

[34] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7478–7498, Jun. 2023.

[35] E. Irmak, "Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework," *Iranian J. Sci. Technol., Trans. Electr. Eng.*, vol. 45, no. 3, pp. 1015–1036, Sep. 2021.

[36] T. Sadad, A. Rehman, A. Munir, T. Saba, U. Tariq, N. Ayesha, and R. Abbasi, "Brain tumor detection and multi-classification using advanced deep learning techniques," *Microsc. Res. Technique*, vol. 84, no. 6, pp. 1296–1308, Jan. 2021.

[37] M. Ben Naceur, M. Akil, R. Saouli, and R. Kachouri, "Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101692.

[38] K. Muhammad, S. Khan, J. D. Ser, and V. H. C. D. Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 507–522, Feb. 2021.

[39] S. Kumar and D. P. Mankame, "Optimization driven deep convolution neural network for brain tumor classification," *Biocybernetics Biomed. Eng.*, vol. 40, no. 3, pp. 1190–1204, Jul. 2020.

[40] A. A. Asiri, A. Shaf, T. Ali, M. Aamir, M. Irfan, and S. Alqahtani, "Enhancing brain tumor diagnosis: An optimized CNN hyperparameter model for improved accuracy and reliability," *PeerJ Comput. Sci.*, vol. 10, p. e1878, Mar. 2024.

[41] A. A. Asiri, A. Shaf, T. Ali, M. A. Pasha, M. Aamir, M. Irfan, S. Alqahtani, A. J. Alghamdi, A. H. Alghamdi, A. F. A. Alshamrani, M. Alelyani, and S. Alamri, "Advancing brain tumor classification through fine-tuned vision transformers: A comparative study of pre-trained models," *Sensors*, vol. 23, no. 18, p. 7913, Sep. 2023.

[42] A. A. Asiri, M. Aamir, T. Ali, A. Shaf, M. Irfan, K. M. Mehdar, S. M. Alqhtani, A. H. Alghamdi, A. F. A. Alshamrani, and O. M. Alshehri, "Next-gen brain tumor classification: Pioneering with deep learning and fine-tuned conditional generative adversarial networks," *PeerJ Comput. Sci.*, vol. 9, p. e1667, Nov. 2023.

[43] E. Gürsoy and Y. Kaya, "Brain-GCN-net: Graph-convolutional neural network for brain tumor identification," *Comput. Biol. Med.*, vol. 180, Sep. 2024, Art. no. 108971.

[44] A. Khuzaim Alzahrani, A. A. Alsheikhy, T. Shawly, A. S. Azzahrani, and A. I. AbuEid, "Amyotrophic lateral sclerosis prediction framework using a multi-level encoders–decoders-based ensemble architecture technology," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 2, Feb. 2024, Art. no. 101960.

[45] T.-W. Wang, Y.-C. Shiao, J.-S. Hong, W.-K. Lee, M.-S. Hsu, H.-M. Cheng, H.-C. Yang, C.-C. Lee, H.-C. Pan, W. C. You, J.-F. Lirng, W.-Y. Guo, and Y.-T. Wu, "Artificial intelligence detection and segmentation models: A systematic review and meta-analysis of brain tumors in magnetic resonance imaging," *Mayo Clinic Proceedings: Digit. Health*, vol. 2, no. 1, pp. 75–91, Mar. 2024.

[46] E. Moya-Sáez, R. De Luis-García, L. Nunez-Gonzalez, C. Alberola-López, and J. A. Hernández-Tamames, "Brain tumor enhancement prediction from pre-contrast conventional weighted images using synthetic multiparametric mapping and generative artificial intelligence," *Quant. Imag. Med. Surgery*, vol. 15, no. 1, pp. 42–54, Jan. 2025.

[47] M. Sravani, S. Aparna, J. Sabarinath, and Y. Kakarla, "Enhancing brain tumor diagnosis with generative adversarial networks," in *Proc. 14th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2024, pp. 846–851.

ANEES TARIQ received the B.Sc. degree in computer science from COMSATS University Islamabad, Wah Campus. He is currently pursuing the master's degree with UET Taxila. He was a Graduate Assistant with UET Taxila. His current research interests include image processing, medical image analysis, machine learning, and computer vision.

MUHAMMAD MUNWAR IQBAL received the M.Sc. degree in computer science from the University of the Punjab, Lahore, the M.S. degree in computer science from the COMSATS Institute of Information Technology, Lahore, in 2011, and the Doctor of Philosophy, under Dr. Yasir Saleem, an Associate Professor at the Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan. He is currently an Assistant Professor with the Department of Computer Science, University of Engineering and Technology at Taxila, Taxila, Pakistan. He has authored and co-authored journal and conference papers in computer science at national and international levels. His research interests include machine learning, databases, semantic web, eLearning, and artificial intelligence. He bears responsibilities at the Computer Science Department, as the Director of the Academic Cell, the Head of the Semester Committee, Scholarship Committee, Exam Scrutiny Committee, and Curriculum Revision Committee, HEC Laptop Scheme (focal person), a security focal person, and an advisor COMPTECH Society, BSCS 2016 Session, and Prospectus Amendment Committee.

MUHAMMAD JAVED IQBAL received the M.Sc. degree in computer science from the University of Agriculture, Faisalabad, Pakistan, in 2001, the M.S./M.Phil. degree in computer science from International Islamic University Islamabad, Pakistan, in 2008, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2015. He is currently a HEC-approved Ph.D. Supervisor and an Assistant Professor with the Computer Science Department, University of Engineering and Technology at Taxila, Pakistan. His work has been published in several international publications, including journals, book chapters, and conferences. He is also a reviewer of renowned national and international journals and conferences.

IFTIKHAR AHMAD (Member, IEEE) received the B.Sc. degree from The Islamia University of Bahawalpur, Bahawalpur, Pakistan, in 1999, the M.Sc. degree in computer science from the University of Agriculture, Faisalabad, Pakistan, in 2001, the M.S./M.Phil. degree in computer science from the COMSATS Institute of Information Technology, Abbottabad, Pakistan, in 2007, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2011. He has been a Faculty Member and a Research Supervisor at various universities, since 2001. He is currently a Faculty Member with the Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University. He has been involved in several funded projects as PI and Co-PI. He has published several papers in reputed journals and conferences. He is also a member of several scientific and professional bodies.