JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 1, JANUARY, 2026, Pp. 127 – 142
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Comparative Analysis of Hybrid CNN-ViT and CNN for Brain Tumor Classification

Ahmad Fauzi [(1)*], Achmad Lutfi Fuadi [(2)], Agus Heri Yunial [(3)]
Departemen Teknik Informatika, Universitas Pamulang, Tangerang, Indonesia
e-mail : {dosen02621,dosen02524,dosen02525}@unpam.ac.id.
* Corresponding author.

This article was submitted on 5 December 2025, revised on 13 January 2026, accepted on 13 January 2026, and published on 25 January 2026.

Abstract

The automated categorization of brain cancers from MRI is essential for improving diagnostic precision. Traditional Convolutional Neural Networks (CNNs) are proficient in local feature extraction but are constrained in their ability to capture long-range spatial relationships, hence impairing performance on intricate malignancies. We propose a hybrid parallel architecture that merges a CNN with a Vision Transformer (ViT) to combine local and global feature modeling. We assessed our dual-branch model in comparison to a conventional CNN baseline using a curated dataset of 15,000 MRI images categorized into three classes: glioma, meningioma, and pituitary. The hybrid model exhibited enhanced performance, attaining 98.40% accuracy and 0.0783 loss, in contrast to the baseline's 97.40% accuracy and 0.1187 loss. The substantial decrease in misclassifications was validated by additional metrics, such as enhanced recall for the meningioma category. The integration of local and global variables produces a more precise, stable, and generalizable classification framework, demonstrating significant potential as a basis for dependable AI-driven Clinical Decision Support Systems (CDSS) in neuroradiology.

Keywords: Artificial Intelligence, Convolutional Neural Network, Machine Learning, Medical Image Analysis, Vision Transformer

Abstrak

Kategorisasi otomatis kanker otak dari citra MRI sangat penting untuk meningkatkan ketepatan diagnosis. Jaringan Syaraf Konvolusional (CNN) tradisional memiliki kemampuan yang baik dalam mengekstraksi fitur lokal, tetapi terbatas dalam menangkap hubungan spasial jangka panjang, sehingga mengurangi kinerjanya pada kasus keganasan yang kompleks. Kami mengusulkan arsitektur hibrida paralel yang menggabungkan CNN dengan Vision Transformer (ViT) untuk memadukan pemodelan fitur lokal dan global. Model dua cabang ini dievaluasi dan dibandingkan dengan model CNN konvensional menggunakan dataset terkurasi yang berisi 15.000 citra MRI yang diklasifikasikan ke dalam tiga kelas: glioma, meningioma, dan pituitary. Model hibrida menunjukkan peningkatan kinerja yang signifikan, mencapai akurasi 98,40% dan loss 0,0783, dibandingkan dengan model dasar (baseline) yang memiliki akurasi 97,40% dan loss 0,1187. Penurunan besar dalam kesalahan klasifikasi ini divalidasi melalui metrik tambahan, termasuk peningkatan recall untuk kategori meningioma. Integrasi antara variabel lokal dan global menghasilkan kerangka klasifikasi yang lebih akurat, stabil, dan dapat digeneralisasi dengan baik, menunjukkan potensi besar sebagai dasar bagi Sistem Pendukung Keputusan Klinis (Clinical Decision Support Systems/CDSS) berbasis AI yang andal di bidang neuroradiologi.Abstrak dalam bahasa Indonesia ditulis dengan pola yang sama dengan abstrak dalam Bahasa Inggris hanya tidak perlu dimiringkan.

Kata Kunci: Kecerdasan Buatan, Convolutional Neural Network, Pembelajaran Mesin, Analisis Citra Medis

1. INTRODUCTION

The classification of brain tumors using Magnetic Resonance Imaging (MRI) is a crucial process in the development of efficient diagnostic and treatment strategies in the field of neuroradiology (Aggarwal et al., 2023; Fujima et al., 2023a). MRI is the preferred method for visualizing brain tissue due to its excellent soft tissue contrast, allowing for precise diagnosis of disease disorders (Senan et al., 2022). Manual interpretation of high-dimensional MRI scans is sometimes hampered by their labor-intensive nature and considerable inter-observer variability, leading to diagnostic ambiguity (Alanazi et al., 2022; Omer, 2024). Therefore, advances in automation techniques for image processing are becoming increasingly important to improve current clinical practice (Dai et al., 2025).

The development of deep learning technology has triggered a paradigm shift in medical image analysis by presenting an innovative methodology for image classification (Balamurugan & Gnanamanoharan, 2023; Xie, 2023). Convolutional Neural Network (CNN) is a fundamental architecture in this field, highly adept at hierarchical feature extraction and capable of mimicking the diagnostic accuracy of humans (Jamali et al., 2024; Liu et al., 2024; Parulian, 2025). Nevertheless, CNN faces significant obstacles related to intrinsic local bias. This limits its ability to understand long-term spatial relationships, which is something important for distinguishing tumor subtypes that are visually similar but histologically different (Hasan et al., 2025; Touvron et al., 2022).

To overcome these limitations, Vision Transformer (ViT) emerged as a new method that breaks down images into a series of patches and uses self-attention mechanisms to efficiently represent global spatial dependencies (Ruthven et al., 2023; Wibowo et al., 2025). ViT shows potential in better interpretability and prediction consistency between patches than classic CNN designs; nevertheless, ViT is characterized by enormous data needs and weaker local bias (Khatun et al., 2025).

Given the strengths and weaknesses of both architectures, CNN–ViT hybrid research has been on the rise. This hybrid model combines CNN's advantages in local feature extraction and ViT's capabilities in modeling global spatial relationships (Alayón et al., 2023; Emara et al., 2025; Zhang et al., 2023). This combination has been shown to improve classification accuracy compared to using CNN or ViT separately (Touvron et al., 2022). In the context of brain tumor classification, recent research shows that the CNN–ViT hybrid model provides more precise, more reliable, and more clinically interpretable predictions (Bukhari, 2024; Liu et al., 2023; Murugesan et al., 2025; Yu et al., 2024). However, there is still a significant research gap. Most hybrid studies today use sequential pipelines or partial integrations that have not fully leveraged the synergies between local CNN extraction and the global context of ViT. In addition, comparative evaluation of a robust CNN baseline with a uniform experimental setting is still very limited.

Based on the above observations, this study does not aim to introduce a new CNN-ViT architecture. Instead, it focuses on a controlled comparative evaluation between conventional CNNs and parallel CNN-ViT hybrid models under identical experimental conditions. Although the dual-flow CNN-ViT architecture has been explored in previous studies, there is still limited empirical evidence regarding its actual performance gains when compared to robust CNN baselines using the same dataset, preprocessing pipeline, and training protocols. Therefore, this work emphasizes performance comparisons, resiliency analysis, and computational trade-offs rather than architectural novelty.

2. METHODS

2.1 Dataset Acquisition and Characterization

This study uses the public brain MRI dataset from the Multi Cancer Dataset in the Kaggle repository (Naren, 2024). From this dataset, a subset of 15,000 T1-weighted axial images was selected with a balanced distribution among three categories of histologically confirmed tumors: glioma, meningioma, and pituitary tumors, each of 5,000 images. The selection of this subset is based on the need to ensure class balance, which is critical for the stability of the training and evaluation of multi-class classification models.

The dataset was then stratified into three parts: training (n = 9,600; 64%), validation (n = 2,400; 16%), and testing (n = 3,000; 20%). These divisions do not overlap to guarantee that the model is tested on data that has not been seen at all, thus realistically measuring generalization capabilities. This sharing strategy also follows best practices in the evaluation of deep learning models on medical data to avoid overfitting bias.

It is important to note that the Multi Cancer Dataset does not provide explicit patient-level identifiers. As a result, the dataset separation strategy is carried out at the image level using cascading random separation. As a result, the study could not fully guarantee that slices originating from the same patient did not appear in different subsets (training, validation, and testing). These limitations are inherent in the structure of the dataset and are recognized as a potential source of data leaks, which have also been reported in other studies using the same dataset. A visualization of the morphological characteristics of the three tumor types (glioma, meningioma, pituitary) is presented in Figure 1, which provides visual context and helps the reader understand the anatomical differences between the classes.

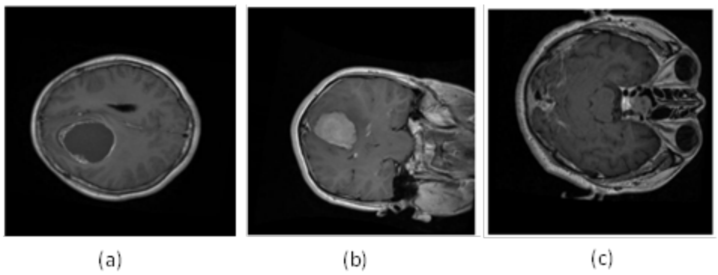

Figure 1 Characteristics of brain tumors: (a) glioma; (b) meningioma; (c) pituitary tumor

2.2 Image Data Pre-processing

A systematic picture pre-processing pipeline was established to guarantee that all model inputs met adequate quality and uniformity standards, thus facilitating an efficient and effective training process (Goyal et al., 2025). Initially, all photos were subjected to pixel intensity equalization with the use of a rescaling factor of 1/255. This approach normalizes pixel values to a range of 0 to 1 (Fujima et al., 2023b), hence improving data uniformity and diminishing model complexity (Goyal et al., 2025). This normalization is essential for expediting model convergence and averting high-intensity pixel values from disrupting the learning process (Goyal et al., 2025). Normalization is formulated as Eq. (1).

$$I_{norm}(x,y) = \frac{I(x,y)}{255} \quad (1)$$

Additionally, all images were scaled to a consistent dimension of 224 × 224 pixels to conform to the input specifications of the CNN architectures, a common procedure in modern image processing applications (Goyal et al., 2025). The data was processed in groups of 32 pictures. In the multi-class classification challenge, labels were one-hot encoded utilizing the 'categorical' mode, a commonly employed method that guarantees a suitable numerical representation for the loss function (Goyal et al., 2025). The division of training and validation data was managed automatically by configuring a validation_split option, which designated 20% of the training data as the validation set. This method enhances data usage and enables more precise evaluation of models during training (Goyal et al., 2025).

It is essential to emphasize that no data augmentation was implemented on the testing set; the photos underwent just normalization. This methodology guarantees that the ultimate model assessment is conducted on unblemished, undistorted data, yielding an accurate evaluation of its performance on novel instances (Goyal et al., 2025). To ensure evaluation consistency, data shuffling was deactivated during the test set processing. This process is crucial for preserving the stability of evaluation metrics and guaranteeing reproducible, reliable outcomes (Ali et al., 2025; Goyal et al., 2025).

2.3 Data Augmentation Strategy

A systematic data augmentation method was exclusively applied to the training images to boost the models' generalization capabilities and extend the distributional representation of the training data (Krichen, 2023). The augmentation process comprised various geometric modifications: random rotations of up to 20 degrees, random horizontal and vertical shifts of up to 20% of the image dimensions, shear transformations, and random zooming (Keng & Merz, 2024; Krichen, 2023). Furthermore, horizontal flipping was utilized to create variations in object orientation, a method recognized for enhancing model robustness against input variations (Dragan et al., 2023). The augmentations were executed in real-time during the training phase via the ImageDataGenerator class. This method eliminates the necessity for explicit storage of augmented images, therefore guaranteeing memory and computational efficiency within the data processing pipeline (Checcucci et al., 2023).

2.4 Architectural Design

This study encompassed the design, execution, and comparative assessment of two separate deep learning systems. The first model is a standard Convolutional Neural Network (CNN), functioning as the experimental baseline to set a performance benchmark. The second is our suggested parallel hybrid CNN–ViT architecture, designed to address the intrinsic constraints of the baseline model.

Both architectures underwent comparable pre-processing pipelines, data partitioning systems, and training hyperparameters to guarantee a fair and rigorous comparison. This method isolates architectural design as the principal variable affecting performance outcomes. The comprehensive methodological architecture of this investigation, encompassing data collection to final model evaluation, is visually encapsulated in Figure 2.

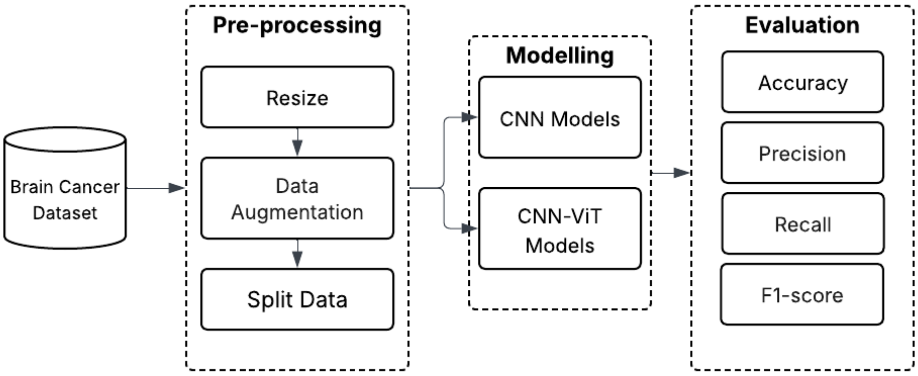

Figure 2 Research Methodology

2.5 Baseline CNN Architecture

The foundational Convolutional Neural Network (CNN) model was developed with the Keras Sequential API, a framework chosen for its efficacy and resilience in prototyping and constructing deep learning models (Pumperla & Cahall, 2022). The architecture, specified in Table 1, consists of two main components: a feature extraction backbone and a classification head. The feature extraction backbone is engineered to handle RGB picture inputs measuring 224 × 224 pixels and comprises five consecutive convolutional blocks. Each block consists of a Conv2D layer for feature extraction, succeeded by BatchNormalization to enhance learning stability and expedite convergence, and a MaxPooling2D layer for spatial down-sampling. The quantity of filters escalates systematically over these blocks (e.g., 32, 64, 128), facilitating the model's capacity for hierarchical feature extraction, hence capturing patterns of escalating complexity from basic edges to elaborate textures.

Subsequent to the convolutional backbone, the resultant feature maps are flattened into a vector and transmitted to the classification head. This skull consists of three fully connected (Dense) layers. A Dropout layer with a rate of 0.5 is employed as a regularization technique to mitigate overfitting. The architecture concludes with a Dense layer employing a Softmax activation function, producing a probability distribution among the three tumor types, which is calculated using the formula in Eq. (2).

$$P(y = i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{2}$$

Table 1 Architectural Specifications of the Baseline CNN Model.

Layer	Output Size	Number of Filters/Unit	Kernel Size	Pooling
Conv2D + Batch Normalization	224×224	32	3×3	–
MaxPooling2D	112×112	–	–	2×2
Conv2D + Batch Normalization	112×112	64	3×3	–
MaxPooling2D	56×56	–	–	2×2
Conv2D + Batch Normalization	56×56	128	3×3	–
MaxPooling2D	28×28	–	–	2×2
Conv2D + Batch Normalization	28×28	256	3×3	–
MaxPooling2D	14×14	–	–	2×2
Conv2D + Batch Normalization	14×14	512	3×3	–
MaxPooling2D	7×7	–	–	2×2

2.6 Hybrid CNN–ViT Architecture

The suggested hybrid model was developed utilizing the Keras Functional API, a framework that offers the necessary flexibility for designing intricate, non-linear network topologies, including a parallel dual-stream architecture (Ali et al., 2025). Figure 3 demonstrates that this architecture is engineered to concurrently process an input image via two separate yet parallel pathways: a CNN-based feature extractor for local patterns and a Vision Transformer (ViT) backbone for global context modeling. This synergistic methodology, demonstrated to be beneficial in medical image analysis (Emara et al., 2025; Yu et al., 2024), seeks to develop a more nuanced and comprehensive feature representation for the classification of brain tumors (Kim et al., 2025; Tummala et al., 2022).

The initial stream, the CNN component, operates as a specialized local feature extractor. It utilizes the identical foundational architecture outlined in the preceding section, with five consecutive convolutional blocks (Conv2D, BatchNormalization, MaxPooling2D). This branch processes the 224×224 RGB input image and is chiefly tasked with capturing intricate spatial hierarchies, including textures and morphological characteristics, essential for local tumor classification. The second stream operates concurrently, employing a pre-trained Vision Transformer model, namely the ViT-Base-Patch16-224 version from the Hugging Face library. This ViT backbone analyzes the identical normalized input by initially segmenting it into a sequence of 16×16 patches. The patches are subsequently linearly embedded and integrated using positional encodings (Wu et al., 2020) to preserve spatial information, as shown in Eq. (3). In this formulation, E is an embedding matrix, E_{pos} is positional encoding, and N is the number of patches. The generated sequence of tokens is processed by the Transformer's self-attention layers using the relation described in Eq. (4), enabling the model to grasp long-range dependencies and overarching contextual linkages throughout the entire image (Deng et al., 2009). The model was initialized with weights pre-trained on the ImageNet dataset to utilize transferred information and enhance training stability.

$$z_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos} \tag{3}$$

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

$$F_{fusion} = [F_{CNN} \parallel F_{ViT}] \tag{5}$$

The feature vectors obtained from the terminal layers of both the CNN and ViT branches are subsequently amalgamated. The fusion is accomplished by a Concatenation layer that integrates the local and global feature representations into a singular, cohesive vector, as expressed in Eq. (5). The concatenated vector is then processed by a classification head consisting of many Dense layers, which are regularized by L2 regularization and Dropout. The architecture concludes with a Dense layer utilizing a Softmax activation function to generate probability scores for the three tumor classifications: glioma, meningioma, and pituitary, in accordance with known methodologies (Touvron et al., 2022).

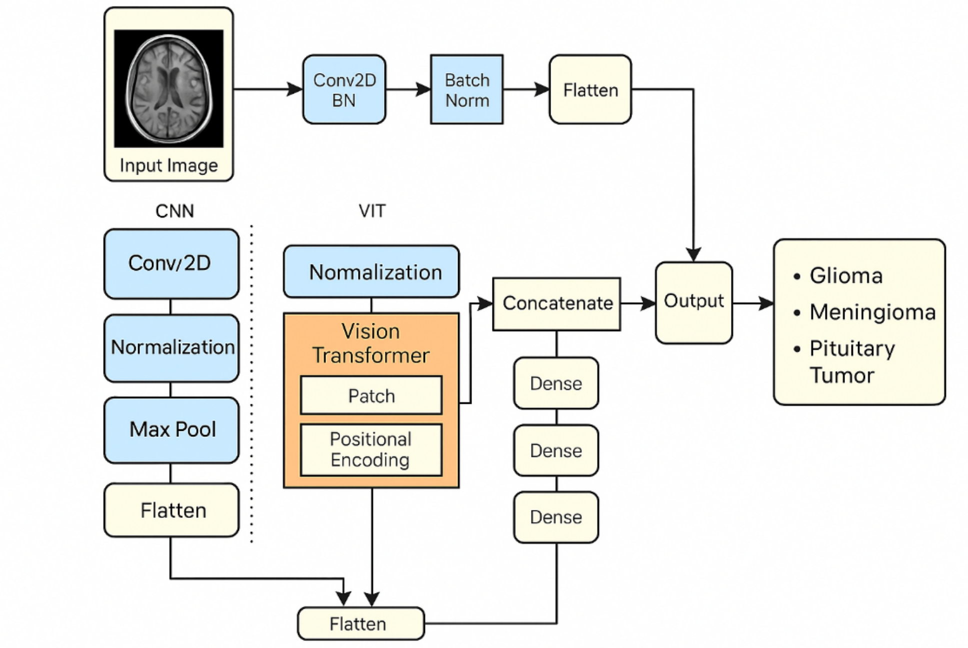

Figure 3 Visualization of the CNN–ViT hybrid architecture

Figure 3 is the architectural diagram of the proposed parallel dual-stream hybrid CNN-ViT model. The workflow commences with a 224×224 pixel MRI input, concurrently processed by two parallel streams. The CNN component focuses on extracting local spatial characteristics, whereas the ViT backbone encompasses global contextual links. The feature vectors generated by each stream are subsequently concatenated and forwarded via a classification head composed of dense layers. The final softmax output layer produces the categorization probabilities for the three tumor categories: glioma, meningioma, and pituitary.

2.7 Vision Transformer Fine-Tuning Strategy

In this study, the Vision Transformer backbone (ViT-Base-Patch16-224) was fully fine-tuned without freezing any of its layers. The model was initialized using ImageNet-21k pretrained weights and trained jointly with the CNN branch in an end-to-end manner. Feature extraction from the ViT branch was performed using the [CLS] token obtained from the last hidden state, which represents the global image-level representation. The same learning rate was applied to both CNN and ViT components to maintain a unified optimization strategy.

2.8 Training and Evaluation Protocol

The evaluation process was created to objectively analyze and compare the efficacy of the two proposed architectures in classifying three types of brain cancers using MRI scans (Murugesan et al., 2025). The traditional CNN architecture functioned as the performance baseline for evaluating the effectiveness of the suggested hybrid CNN–ViT model (Ullah et al., 2023). A range of callback mechanisms was utilized to enhance the training process and mitigate overfitting. This incorporated EarlyStopping with a patience of five epochs, which automatically terminated training if the validation loss did not improve. The ModelCheckpoint callback was set up to preserve solely the model weights that achieved optimal performance on the validation set. Furthermore, ReduceLROnPlateau was employed to dynamically modify the learning rate, decreasing it when progress in learning plateaued. This amalgamation of callbacks is an established technique for augmenting training stability and promoting model generalization (Murugesan et al., 2025). Both models underwent training for a maximum of 50 epochs, with performance evaluated across the training, validation, and testing datasets (Ullah et al., 2023).

A confusion matrix (Eq. (10)) was produced for a detailed performance evaluation on the hold-out test set. Key classification metrics, specifically precision (Eq. (6)), recall (Eq. (7)), and the F1-score (Eq. (8)), were derived for each tumor type from this matrix (Yohannes & Al Rivan, 2022). The assessment findings from the baseline CNN established a vital benchmark for statistically assessing the performance improvements attained through the synergistic integration of CNN and Vision Transformer components in the proposed hybrid architecture (Ullah et al., 2023). Accuracy is mathematically defined as in Eq. (9).

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1Score = 2\ \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Confusion\,Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{10}$$

2.9 Statistical Significance Analysis

To assess whether the performance difference between the baseline CNN and the hybrid CNN–ViT model was statistically significant, a McNemar test was conducted using paired predictions on the same test set. This test is suitable for paired classification comparisons, as it focuses on prediction disagreements rather than overall accuracy values. The McNemar statistic is computed using Eq. (11).

$$X^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{11}$$

3. RESULTS AND DISCUSSION

Two model architectures, a baseline Convolutional Neural Network (CNN) and the proposed hybrid CNN–ViT, were trained and evaluated for a three-class brain tumor classification task (glioma, meningioma, and pituitary tumor). Each model underwent training for a maximum of 50 epochs using 224×224 pixel RGB-formatted MRI scans. Performance was measured on three distinct data subsets: training (9,600 images), validation (2,400 images), and testing (3,000 images). All experiments were executed on the Kaggle Notebook platform, employing dual T4 GPU accelerators to ensure computational efficiency and a uniform training environment.

3.1 Performance of the Baseline CNN Model

The baseline CNN model attained a training accuracy of 99.53%, a validation accuracy of 98.75%, and a final test accuracy of 97.40%. The loss on the test set was noted at 0.1187. Figure 4 displays the training and validation accuracy and loss curves. These figures demonstrate a consistent convergence tendency, despite slight oscillations noted during the initial training epochs. Subsequent examination of the classification metrics indicates a macro-averaged precision of 0.98, a recall of 0.97, and an F1-score of 0.97. The confusion matrix, depicted in Figure 5, reveals that misclassifications primarily transpired between the meningioma and pituitary tumor categories. A total of 78 out of 3,000 test photos were inaccurately classified.

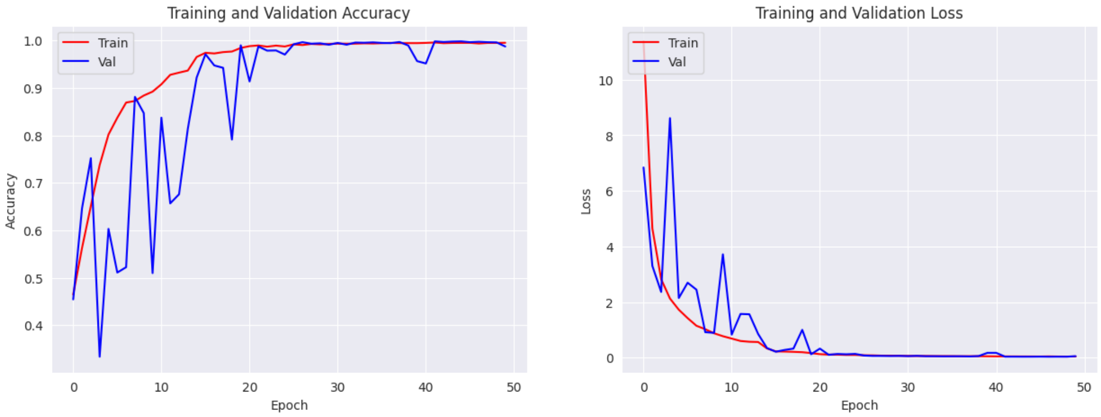

Figure 4 Training & Validation Accuracy

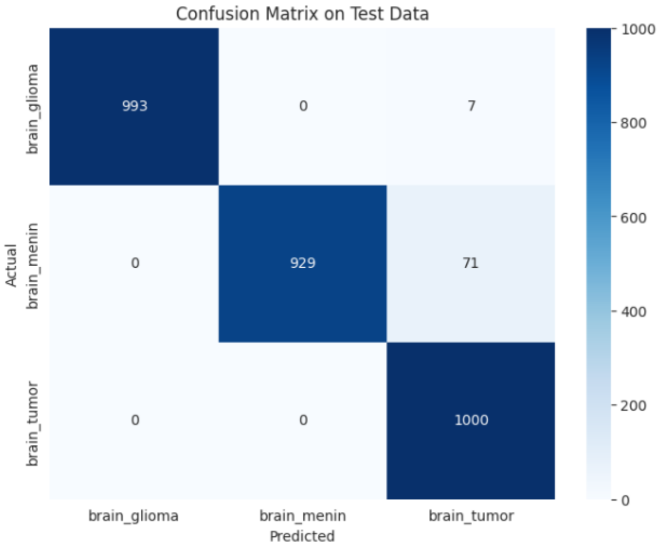

Figure 5 Confusion Matrix Results Baseline CNN Model

The baseline CNN features around 14.7 million trainable parameters and an average training duration of about 126 seconds per epoch. Although these data indicate significant computational efficiency, the model's previously acknowledged shortcoming in modeling global spatial interdependence is demonstrated by the observed discrepancies in per-class recall rates. Table 2 presents a detailed breakdown of the model's complexity and performance measures.

Table 2 Summary of Complexity and Training Performance for the Baseline CNN Model.

Epoch	Akurasi Train	Loss Train	Akurasi Val	Loss Val	Learning Rate	Waktu/Epoch (s)
1	0.4396	14.6324	0.3333	8.6493	5.0×10^{-4}	211
5	0.7988	1.8159	0.6988	1.9758	5.0×10^{-4}	173
10	0.9269	0.6411	0.8929	0.6110	1.5×10^{-4}	173
14	0.9559	0.3708	0.9717	0.2795	4.5×10^{-5}	172
21	0.9803	0.1865	0.9879	0.1548	4.5×10^{-5}	174
28	0.9833	0.1316	0.9925	0.1038	1.35×10^{-5}	174
34	0.9893	0.0887	0.9917	0.0783	1.35×10^{-5}	173
40	0.9923	0.0718	0.9937	0.0661	1.0×10^{-5}	173
47	0.9941	0.0616	0.9962	0.0518	1.0×10^{-5}	173
50	0.9941	0.0550	0.9971	0.0458	1.0×10^{-5}	173

3.2 Performance of the Hybrid CNN–ViT Model

The hybrid CNN–ViT architecture was developed to integrate the local feature extraction capabilities of a CNN with the global context modeling of a Vision Transformer. This was accomplished by incorporating a pre-trained ViT-Base-Patch16-224 model from the Hugging Face library into a parallel stream, markedly improving the model's feature representation capability. The hybrid CNN–ViT model, trained in the same configuration as the baseline, attained a training accuracy of 99.41%, a validation accuracy of 99.71%, and a final test accuracy of 98.40%. The test loss was 0.0783, indicating a significant 34% decrease relative to the baseline CNN. The training curves in Figure 7 illustrate that the hybrid model demonstrated a significantly steadier convergence pattern compared to the baseline. The slight variations in the validation curve specifically indicate enhanced generalization ability.

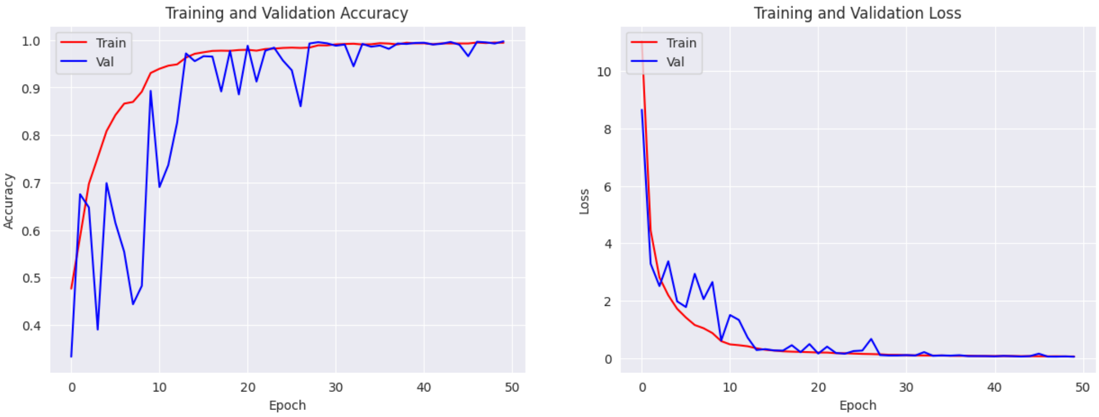

Figure 6 Training & Validation Loss Hybrid CNN-VIT

The hybrid model attained a consistent macro-averaged precision, recall, and F1-score of 0.98, as determined by the assessment of classification measures. The confusion matrix, illustrated in Figure 8, indicates a significant drop in classification errors, with merely 40 misclassified instances out of 3,000 test images, or a 48.7% reduction relative to the baseline. Notable enhancements in accuracy were evident in distinguishing between the meningioma and pituitary tumor categories.

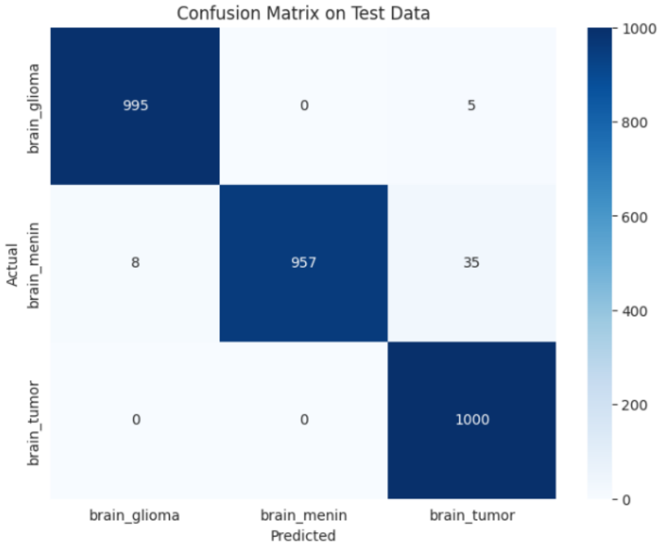

Figure 7 Confusion Matrix Results Hybrid CNN-VIT

Nonetheless, this enhancement in efficiency is coupled with a considerable rise in model complexity. The hybrid CNN–ViT architecture contains around 85 million trainable parameters, primarily attributable to the incorporation of the extensive ViT backbone. The average training duration per epoch thus rose to around 172 seconds, about 36% higher than that of the baseline CNN. Nonetheless, the significant enhancements in accuracy and predictive stability are contended to warrant this computational burden, especially in clinical applications where diagnostic dependability is crucial. Table 3 delineates a summary of the hybrid model's complexity and performance characteristics.

Table 3 Summary of Complexity and Training Performance for the Hybrid CNN–ViT Model.

Epoch	Akurasi Train	Loss Train	Akurasi Val	Loss Val	Learning Rate	Waktu/Epoch (s)
1	0.4396	14.6324	0.3333	8.6493	5.0×10^{-4}	211
5	0.7988	1.8159	0.6988	1.9758	5.0×10^{-4}	173
10	0.9269	0.6411	0.8929	0.6110	1.5×10^{-4}	173
14	0.9559	0.3708	0.9717	0.2795	4.5×10^{-5}	172
21	0.9803	0.1865	0.9879	0.1548	4.5×10^{-5}	174
28	0.9833	0.1316	0.9925	0.1038	1.35×10^{-5}	174
34	0.9893	0.0887	0.9917	0.0783	1.35×10^{-5}	173
40	0.9923	0.0718	0.9937	0.0661	1.0×10^{-5}	173
47	0.9941	0.0616	0.9962	0.0518	1.0×10^{-5}	173
50	0.9941	0.0550	0.9971	0.0458	1.0×10^{-5}	173

3.3 Statistical Significance Analysis

To evaluate the statistical significance of the performance differences between the CNN model and the CNN–ViT hybrid, a McNemar test was performed based on prediction pairs on the same test data with reference to Equation (9). Based on the confusion matrix, the CNN model produced a total of 78 misclassifications, dominated by meningioma errors that were misclassified as tumors (71 cases), while the CNN–ViT hybrid model resulted in 48 errors with a lower error distribution in the same class. Prediction pair analysis showed that the CNN–ViT hybrid model managed to correct 38 errors previously made by CNN ($b = 38$), while generating only 8 new errors that did not appear in the CNN model ($c = 8$). By substituting the values b and c into Equation (9), the statistical value χ^2 is obtained as 18.28 with a degree of freedom of one, which results in a p-value < 0.001. These results show that the performance improvements achieved by the CNN–ViT hybrid model are statistically significant and are not caused by mere random variation.

3.4 Comparative Analysis Baseline CNN and Hybrid CNN–ViT

A thorough comparative analysis was done to evaluate the efficacy of the proposed hybrid architecture between the baseline CNN and the hybrid CNN–ViT models. The comparison included a thorough array of performance parameters, including accuracy, loss, precision, recall, and the F1-score, evaluated throughout the training, validation, and hold-out test sets. In addition to predicted accuracy, the investigation examined critical technical factors, including the overall count of trainable parameters and the average training duration per session. These factors offer a comprehensive assessment of the trade-offs among model efficacy, computing efficiency, and scalability. Table 4 summarizes the findings of this head-to-head comparison, offering a quantitative basis for the ensuing debate.

Table 4 Comparative Summary of Performance and Complexity for Baseline and Hybrid Models.

Evaluation Aspect	Baseline CNN	Hybrid CNN–ViT
Training Accuracy	99.53%	99.41%
Validation Accuracy	98.75%	99,71%
Test Accuracy	97.40%	98.40%
Test Loss	0.1187	0.0783
Precision (average)	0.98	0.98
Recall (average)	0.97	0.98
F1-Score (average)	0.97	0.98
Number of Prediction Errors	78	40
Number of Parameters	±14,7 million	±85 million
Training Time/Epoch	±126 seconds	±172 seconds

This paper presents a thorough assessment of an innovative hybrid CNN–ViT architecture aimed at overcoming ongoing difficulties in the automated categorization of brain tumors using MRI data. The results validate that the suggested hybrid model exhibits a substantial and measurable performance superiority compared to a traditional CNN baseline. This result provides robust empirical support for the theoretical assertion that synergistically integrating the local feature extraction abilities of CNNs with the global context modeling of Vision Transformers constitutes a very effective approach (Ishrak et al., 2025). The subsequent sections will analyze these findings, addressing their theoretical and practical consequences, and will conclude with a summary of the study's shortcomings and potential directions for further research.

Table 5 Comparison with Related Studies on Brain Tumor Classification

No.	Study	Model Architecture	Dataset	Accuracy
1	Tummala et al. (2022)	ViT Ensemble (B/16, B/32, L/16, L/32)	Figshare Brain MRI	98.7%
2	Ullah et al. (2023)	Enhanced CNN (VGG16, VGG19, ResNet101, InceptionV3)	Public Brain MRI	97.0%
3	Ali et al. (2025)	ResNet50	Kaggle Brain MRI	99.88%
4	Emara et al. (2025)	Unified CNN–ViT (HViT-CNN)	Multi-domain MRI (Brain)	98.4%
5	This Study	Parallel CNN–ViT (Dual-Stream)	Multi Cancer MRI	98.4%

The selected works encompass CNN-based, Vision Transformer-based, and hybrid CNN–Transformer architectures to provide a balanced contextual comparison with the proposed method. It is important to note that variations in reported performance may arise from differences in dataset characteristics, network design, and evaluation protocols. Accordingly, this comparison is intended to highlight architectural trends and relative performance rather than to assert absolute superiority.

Based on the comparisons presented in Table 5, it can be observed that both the Vision Transformer, CNN-based approaches, and the CNN–Transformer hybrid architecture have shown competitive performance in the classification of brain tumors based on MRI images. The study by Tummala et al. (2022) emphasizes the power of global representation through the ViT ensemble, while Ali et al. (2025) and Ullah et al. (2023) show that an optimized CNN is still capable of achieving high accuracy on certain datasets. On the other hand, a hybrid approach such as that proposed by Emara et al. (2025) indicates the potential for the integration of local and global features within a single unified framework. In line with these findings, the results of this study show that the parallel CNN–ViT architecture is able to achieve performance comparable to the current approach, while maintaining classification stability and error reduction between classes. The difference in performance achievement between studies is influenced by the variation in the dataset, the number of classes, and the evaluation protocol used, so this comparison is intended to provide an empirical context for the position of this research in the existing research landscape.

3.5 Theoretical Implications

This study's primary conclusion offers strong empirical support for a fundamental theoretical principle in computer vision: the combined integration of local and global feature extractors produces a more effective and comprehensive data representation. Our hybrid model, which combines a CNN's ability to capture intricate local information with a ViT's capability to model extensive spatial dependencies, surpassed the performance of the standalone CNN architecture. The model's capacity to distinguish between physically identical tumor types, including meningioma and pituitary, is particularly clear, as indicated by a nearly 50% decrease in misclassification errors. This outcome provides robust empirical validation for the assertion made by Ishrak et al. (2025), which contends that hybrid models thrive specifically due to their integration of these two complementary feature extraction paradigms. Moreover, the efficacy of this parallel integration substantiates the findings of other scholars that the amalgamation of Transformer-based and convolutional techniques represents a highly promising and theoretically robust avenue for the progression of medical image analysis (Avcı, 2025; Ullah et al., 2023).

3.6 Practical Implications

This research illustrates the feasibility of a high-performance model that reconciles accuracy with deployability. The proposed hybrid CNN-ViT, exhibiting a test accuracy of 98.40% and a consistently elevated F1-score of 0.98, serves as a more dependable instrument for prospective clinical application than the baseline. Although this performance enhancement entails greater computing complexity, the trade-off is probably warranted in a clinical setting where diagnostic reliability is essential, a perspective aligned with the findings of (Tabassum & Nunavath, 2024).

Furthermore, in comparison to previous high-precision models, our methodology presents a unique benefit. It circumvents the intricate, resource-demanding ensembling methods necessitated by certain models (Ma et al., 2025) and the operational difficulties associated with multi-stage systems (So-yun Park et al., 2025), offering a more efficient, end-to-end solution. Thus, the architecture established in this research can provide a solid basis for the forthcoming generation of AI-enhanced Clinical Decision Support Systems (CDSS). This corroborates the assertion by (Hossain et al., 2025) that these hybrid models are positioned to catalyze substantial, concrete advancements in healthcare and medical diagnostics.

3.7 Limitations and Future Directions

Notwithstanding the encouraging findings, this study possesses multiple limitations that delineate explicit opportunities for subsequent research. A significant limitation is the model's considerable computational burden, a challenge observed in other high-performance hybrid methodologies (Park et al., 2025). Consequently. Consequently, future research should prioritize model optimization strategies (e.g., pruning, quantization) and validation across varied, multi-institutional datasets to guarantee clinical reliability. Moreover, the model's therapeutic value could be substantially enhanced by broadening its application to more specific tasks, such as detailed glioma grading or semantic segmentation. Simultaneously, the implementation of explainability (XAI) methodologies is essential for clarifying the model's decision-making process. This will enhance its credibility for clinical adoption, a vital element for implementing such models in practice, as indicated by (Chibuike & Yang, 2024). Addressing these issues will be essential for actualizing the complete potential of hybrid models in practical diagnostic applications (Hossain et al., 2025).

4. CONCLUSIONS

This study introduced and validated an advanced computational method for brain tumor classification by a comparative examination of a conventional Convolutional Neural Network (CNN) and an innovative parallel hybrid CNN–Vision Transformer (ViT) architecture. The main goal was to improve classification accuracy by combining the local feature extraction abilities of CNNs with the global spatial representation strengths of ViTs. The experimental findings indicated that the suggested hybrid CNN–ViT architecture consistently surpassed the traditional CNN baseline across all primary assessment parameters. The hybrid model attained a final test accuracy of 98.40%, exceeding the baseline CNN's 97.40%, and demonstrated enhanced validation stability along with a significant decrease in inter-class confusion.

This research demonstrates that incorporating a ViT into a parallel CNN framework markedly boosts spatial modeling and augments the model's generalization skills on novel data. This method increases computational complexity, although the significant improvements in predicted accuracy and reliability provide a strong rationale for its use in critical clinical settings. This study significantly contributes to the advancement of deep learning-based diagnostic assistance systems in oncological radiology. It also establishes a novel trajectory for the investigation of parallel architectures that adeptly integrate local and global domains for enhanced medical image representation.

REFERENCES

Aggarwal, K., Manso Jimeno, M., Ravi, K. S., Gonzalez, G., & Geethanath, S. (2023). Developing and Deploying Deep Learning Models in Brain Magnetic Resonance Imaging: A Review. *NMR in Biomedicine*, *36*(12), Article ID: e5014. https://doi.org/10.1002/nbm.5014

Alanazi, M. F., Ali, M. U., Hussain, S. J., Zafar, A., Mohatram, M., Irfan, M., AlRuwaili, R., Alruwaili, M., Ali, N. H., & Albarrak, A. M. (2022). Brain Tumor/Mass Classification Framework Using Magnetic-Resonance-Imaging-Based Isolated and Developed Transfer Deep-Learning Model. *Sensors*, *22*(1), Article ID: 372. https://doi.org/10.3390/s22010372

Alayón, S., Hernández, J., Fumero, F. J., Sigut, J. F., & Díaz-Alemán, T. (2023). Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. *Applied Sciences*, *13*(23), Article ID: 12722. https://doi.org/10.3390/app132312722

Ali, R. R., Yaacob, N. M., Alqaryouti, M. H., Sadeq, A. E., Doheir, M., Iqtait, M., Rachmawanto, E. H., Sari, C. A., & Yaacob, S. S. (2025). Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50. *Diagnostics*, *15*(5), Article ID: 624. https://doi.org/10.3390/diagnostics15050624

Avcı, D. (2025). A New Pes Planus Automatic Diagnosis Method: ViT-OELM Hybrid Modeling. *Diagnostics*, *15*(7), Article ID: 867. https://doi.org/10.3390/diagnostics15070867

Balamurugan, T., & Gnanamanoharan, E. (2023). Brain Tumor Segmentation and Classification Using Hybrid Deep CNN with LuNetClassifier. *Neural Computing and Applications*, *35*(6), 4739–4753. https://doi.org/10.1007/s00521-022-07934-7

Bukhari, M. T. (2024). Efficacy of Lightweight Vision Transformers in Diagnosis of Pneumonia. *International Journal of Clinical Nephrology*, *3*(1). https://doi.org/10.37579/2834-5142/020

Checcucci, E., Piazzolla, P., Marullo, G., Innocente, C., Salerno, F., Ulrich, L., Moos, S., Quarà, A., Volpi, G., Amparore, D., Piramide, F., Turcan, A., Garzena, V., Garino, D., De Cillis, S., Sica, M., Verri, P., Piana, A., Castellino, L., … Porpiglia, F. (2023). Development of Bleeding Artificial Intelligence Detector (BLAIR) System for Robotic Radical Prostatectomy. *Journal of Clinical Medicine*, *12*(23), Article ID: 7355. https://doi.org/10.3390/jcm12237355

Chibuike, O., & Yang, X. (2024). Convolutional Neural Network–Vision Transformer Architecture with Gated Control Mechanism and Multi-Scale Fusion for Enhanced Pulmonary Disease Classification. *Diagnostics*, *14*(24), Article ID: 2790. https://doi.org/10.3390/diagnostics14242790

Dai, Y., Lian, C., Zhang, Z., Gao, J., Lin, F., Li, Z., Wang, Q., Chu, T., Aishanjiang, D., Chen, M., Wang, X., Cheng, G., Huang, R., Dong, J., Zhang, H., & Mao, N. (2025). Development and Validation of a Deep Learning System to Differentiate HER2-Zero, HER2-Low, and HER2-Positive Breast Cancer Based on Dynamic Contrast-Enhanced MRI. *Journal of Magnetic Resonance Imaging*, *61*(5), 2212–2220. https://doi.org/10.1002/jmri.29670

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dragan, P., Merski, M., Wiśniewski, S., Sanmukh, S. G., & Latek, D. (2023). Chemokine Receptors—Structure-Based Virtual Screening Assisted by Machine Learning. *Pharmaceutics*, *15*(2), Article ID: 516. https://doi.org/10.3390/pharmaceutics15020516

Emara, H. M., El-Shafai, W., Soliman, N. F., Algarni, A. D., El-Samie, F. E. A., & Mahmoud, A. A. (2025). A Unified Vision Transformer and Convolutional Neural Network Framework for Multi-Domain Cancer Classification. In *Research Square*. https://doi.org/10.21203/rs.3.rs-6633290/v1

Fujima, N., Kamagata, K., Ueda, D., Fujita, S., Fushimi, Y., Yanagawa, M., Ito, R., Tsuboyama, T., Kawamura, M., Nakaura, T., Yamada, A., Nozaki, T., Fujioka, T., Matsui, Y., Hirata, K., Tatsugami, F., & Naganawa, S. (2023a). Current State of Artificial Intelligence in Clinical Applications for Head and Neck MR Imaging. *Magnetic Resonance in Medical Sciences*, *22*(4), 401–414. https://doi.org/10.2463/mrms.rev.2023-0047

Fujima, N., Kamagata, K., Ueda, D., Fujita, S., Fushimi, Y., Yanagawa, M., Ito, R., Tsuboyama, T., Kawamura, M., Nakaura, T., Yamada, A., Nozaki, T., Fujioka, T., Matsui, Y., Hirata, K., Tatsugami, F., & Naganawa, S. (2023b). Current State of Artificial Intelligence in Clinical Applications for Head and Neck MR Imaging. *Magnetic Resonance in Medical Sciences*, *22*(4), 401–414. https://doi.org/10.2463/mrms.rev.2023-0047

Goyal, A. K., Pandey, M., Singh, D. P., Choudhary, J., & Haripriya, R. (2025). FedContrast: A Contrastive Learning Framework to Mitigate Client Drift Under Statistical Heterogeneity in Federated Multi-Class Soil Classification. In *Research Square*. https://doi.org/10.21203/rs.3.rs-6774369/v1

Hasan, M. A., Haque, F., Sarker, H., Abdullah, R., Roy, T., Taaha, N., Arafat, Y., Patwary, A. K., Ahsan, M., & Haider, J. (2025). Mulberry Leaf Disease Detection by CNN-ViT with XAI Integration. *PLOS One*, *20*(6), Article ID: e0325188. https://doi.org/10.1371/journal.pone.0325188

Hossain, Z., Hossain, M. E., Ahmed, N., Kabir, M. F., & Hossain, I. S. (2025). Evaluating the Performance of Vision Transformers and Convolutional Neural Networks for Hostile Image Detection. *Indonesian Journal of Advanced Research*, *4*(1), 111–130. https://doi.org/10.55927/ijar.v4i1.13681

Ishrak, M. F., Rahman, M. M., Joy, M. I. K., Tamuly, A., Akter, S., Tanim, D. M., Jawar, S., Ahmed, N., & Rahman, M. S. (2025). Vision Transformer Embedded Feature Fusion Model with Pre-Trained Transformers for Keratoconus Disease Classification. *Emerging Science Journal*, *9*(2), 1037–1075. https://doi.org/10.28991/ESJ-2025-09-02-027

Jamali, A., Roy, S. K., Hong, D., Lu, B., & Ghamisi, P. (2024). How to Learn More? Exploring Kolmogorov–Arnold Networks for Hyperspectral Image Classification. *Remote Sensing*, *16*(21), Article ID: 4015. https://doi.org/10.3390/rs16214015

Keng, M., & Merz, K. M. (2024). Eliminating the Deadwood: A Machine Learning Model for CCS Knowledge-Based Conformational Focusing for Lipids. *Journal of Chemical Information and Modeling*, *64*(20), 7864–7872. https://doi.org/10.1021/acs.jcim.4c01051

Khatun, R., Chatterjee, S., Bert, C., Wadepohl, M., Ott, O. J., Semrau, S., Fietkau, R., Nürnberger, A., Gaipl, U. S., & Frey, B. (2025). Complex-Valued Neural Networks to Speed-Up MR Thermometry During Hyperthermia Using Fourier PD and PDUNet. *Scientific Reports*, *15*(1), Article ID: 11765. https://doi.org/10.1038/s41598-025-96071-x

Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis. *Journal of Imaging Informatics in Medicine*, *38*(5), 3248–3262. https://doi.org/10.1007/s10278-024-01322-4

Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers*, *12*(8), Article ID: 151. https://doi.org/10.3390/computers12080151

Liu, S., Yue, W., Guo, Z., & Wang, L. (2024). Multi-Branch CNN and Grouping Cascade Attention for Medical Image Classification. *Scientific Reports*, *14*(1), Article ID: 15013. https://doi.org/10.1038/s41598-024-64982-w

Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y., & Zhou, H. (2023). Deep Learning Based Brain Tumor Segmentation: A Survey. *Complex & Intelligent Systems*, *9*(1), 1001–1026. https://doi.org/10.1007/s40747-022-00815-5

Ma, W., Chen, R., Zhang, K., Wu, S., & Ding, S. (2025). Instruct Where the Model Fails: Generative Data Augmentation via Guided Self-Contrastive Fine-Tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(6), 5991–5999. https://doi.org/10.1609/aaai.v39i6.32640

Murugesan, G., Nagendran, P., & Natarajan, J. (2025). Advancing Brain Tumor Diagnosis: Deep Siamese Convolutional Neural Network as a Superior Model for MRI Classification. *Brain-X*, *3*(2). https://doi.org/10.1002/brx2.70028

Naren, O. S. (2024). *Multi Cancer Dataset*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3415848

Omer, A. A. M. (2024). Image Classification Based on Vision Transformer. *Journal of Computer and Communications*, *12*(04), 49–59. https://doi.org/10.4236/jcc.2024.124005

Park, Saeran, Ayana, G., Wako, B. D., Jeong, K. C., Yoon, S., & Choe, S. (2025). Vision Transformers for Low-Quality Histopathological Images: A Case Study on Squamous Cell Carcinoma Margin Classification. *Diagnostics*, *15*(3), 260. https://doi.org/10.3390/diagnostics15030260

Park, So-yun, Ayana, G., Wako, B. D., Jeong, K. C., Yoon, S.-D., & Choe, S. (2025). Vision Transformers for Low-Quality Histopathological Images: A Case Study on Squamous Cell Carcinoma Margin Classification. *Diagnostics*, *15*(3), Article ID: 260. https://doi.org/10.3390/diagnostics15030260

Parulian, O. S. (2025). Efficient Design and Compression of CNN Models for Rapid Character Recognition. *Jurnal Ilmu Komputer dan Informasi*, *18*(1), 127–140. https://doi.org/10.21609/jiki.v18i1.1443

Pumperla, M., & Cahall, D. (2022). Elephas: Distributed Deep Learning with Keras & Spark. *Journal of Open Source Software*, *7*(80), Article ID: 4073. https://doi.org/10.21105/joss.04073

Ruthven, M., Peplinski, A. M., Adams, D. M., King, A. P., & Miquel, M. E. (2023). Real-Time Speech MRI Datasets with Corresponding Articulator Ground-Truth Segmentations. *Scientific Data*, *10*(1), Article ID: 860. https://doi.org/10.1038/s41597-023-02766-z

Senan, E. M., Jadhav, M. E., Rassem, T. H., Aljaloud, A. S., Mohammed, B. A., & Al-Mekhlafi, Z. G. (2022). Early Diagnosis of Brain Tumour MRI Images Using Hybrid Techniques between Deep and Machine Learning. *Computational and Mathematical Methods in Medicine*, *2022*, 1–17. https://doi.org/10.1155/2022/8330833

Tabassum, I., & Nunavath, V. (2024). A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data. *Applied Sciences*, *14*(24), Article ID: 12007. https://doi.org/10.3390/app142412007

Touvron, H., Cord, M., & Jégou, H. (2022). DeiT III: Revenge of the ViT. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 516–533). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20053-3_30

Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. *Current Oncology*, *29*(10), 7498–7511. https://doi.org/10.3390/curroncol29100590

Ullah, Z., Odeh, A., Khattak, I., & Al Hasan, M. (2023). Enhancement of Pre-Trained Deep Learning Models to Improve Brain Tumor Classification. *Informatica*, *47*(6), 165. https://doi.org/10.31449/inf.v47i6.4645

Wibowo, F. A., Yulianto, T., Malun, N. O., Rionaldy, R., Yasin, V., & Siagian, R. C. (2025). Application of Machine Learning Methods for Classification of Gamma and Hadron Signals in High Energy Particle Detection. *Jurnal Ilmu Komputer dan Informasi*, *18*(2), 181–205. https://doi.org/10.21609/jiki.v18i2.1489

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. http://arxiv.org/abs/2006.03677

Xie, X. (2023). Deep Learning-Based Image Classification of MRI Brain Image. *Theoretical and Natural Science*, *27*(1), 46–51. https://doi.org/10.54254/2753-8818/27/20240670

Yohannes, R., & Al Rivan, M. E. (2022). Klasifikasi Jenis Kanker Kulit Menggunakan CNN-SVM. *Jurnal Algoritme*, *2*(2), 133–144. https://doi.org/10.35957/algoritme.v2i2.2363

Yu, Z., Liu, F., & Li, Y. (2024). scTCA: A Hybrid Transformer-CNN Architecture for Imputation and Denoising of scDNA-seq Data. *Briefings in Bioinformatics*, *25*(6), Article ID: bbae577. https://doi.org/10.1093/bib/bbae577

Zhang, Y., Li, Z., Nan, N., & Wang, X. (2023). TranSegNet: Hybrid CNN-Vision Transformers Encoder for Retina Segmentation of Optical Coherence Tomography. *Life*, *13*(4), Article ID: 976. https://doi.org/10.3390/life13040976