Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# A GENERALIZED HYBRID CNN TRANSFORMER FRAMEWORK WITH BIOINSPIRED AOHHO TUNING FOR HETEROGENEOUS MEDICAL IMAGE CLASSIFICATION APER TITLE HERE

**ANJANEYULU GUDLA[1*],G MAHENDER[2], DR KATAM NAGA LAKSHMAN[3], DR VOOTLA SRIDHAR[4] , PRAVEENA NUTHAKKI[5]**

[1]Assistant Professor, Department of Information Technology, Guru Nanak Institutions Technical Campus, Ibrahimpatnam , RR Dist , Hyderabad, Telangana -501506 , India
[2]Assistant Professor, Department of Computer Science Engineering(Data Science), Verdhaman College of Engineering, Shamshabad, Hyderabad, India
[3]Associate Professor, Department of Computer Science Engineering, MallaReddy(MR) Deemed to be University, Maisammaguda (H)Kompally, Medchal Malkajgiri, Hyderabad-500100, India
[4]Associate Professor, Department of Electronics & Communication Engineering, Pallavi Engineering College, UGC Autonomous,HayathNaga, RR Dist, Hyderabad,Telangana, 501505, India
[5] Department of CSE, Koneru Lakshmaiah Education Foundation, Vadeswaram 522302, Guntur, Indi;

E-mail: [1]anjugudla5@gmail.com, [2]mahender1852@verdhaman.org, [3]katamnagalakshman@gmail.com , [4]sridhar463@gmail.com , praveena.4u@gmail.com

**ABSTRACT**

Clinical diagnosis and treatment planning depend on Magnetic Resonance Imaging (MRI) is an accurate method of classifying brain tumors. However, conventional convolutional neural networks often have difficulty dealing with complex tumor patterns, interclass similarity, and suboptimal parameter settings despite the remarkable Intensification of deep learning techniques in the field of medical image analysis. Using an optimized multi-feature learning strategy and convolutional neural networks, this study develops a hybrid deep learning framework that addresses these challenges. As part of the proposed architecture, spatial representations are extracted using convolutional layers, while far-range contextual dependencies across MRI slices are captured by transformer modules. Moreover, a multi-feature fusion component improves discriminative feature learning by combining spatial and frequency-domain representations. The hyperparameters learning rate, convolutional kernel size, and attention weights are automatically tuned using a hybrid optimization strategy based on Aquila optimizer and Harris Hawks optimization. By combining adaptive feature fusion and evolutionary optimization, classification robustness is improved while overfitting is reduced. A comparison of the proposed framework with conventional CNN-based models and recent hybrid architectures based on benchmark MRI datasets shows superior performance. During training, the accuracy, sensitivity, and F1-score of classifications improved significantly. Moreover, the framework proved its versatility by achieving 97.13% training convergence and 84.50% test accuracy on the PathMNIST histological dataset. A hybrid metaheuristic optimization approach in conjunction with transformer-based attention and multi-feature representation learning provides an effective approach for reliably detecting brain tumors in medical imaging

**Keywords:** *Brain Tumor Classification, Magnetic Resonance Imaging (MRI), Convolutional Neural Networks (CNN), Vision Transformer, Cross-Dataset Validation*

## 1. INTRODUCTION

There is still much to learn about modern medical diagnostics when it comes to brain tumors, Among the most serious neurological disorders in the world, it is crucial to identify brain tumors early and classify them accurately to improve the survival rates of patients and to plan effective treatment. As MRI is capable of capturing detailed anatomical structures and soft tissue variations within the brain, imaging with CT is the most common imaging modality for diagnosing brain tumors. It is however time-consuming process that is susceptible to human error error to manually analyze MRI images by radiologists, especially when tumors exhibit complex shapes, heterogeneous textures, and subtle

variations in intensity. Thus, automated computer-aided diagnostic systems are increasingly being developed so that clinicians can analyze and interpret medical images more efficiently [6], [29].

The performance of automated medical image analysis systems has been significantly improved by recent advances in deep learning. The ability of convolutional neural networks (CNNs) to learn hierarchical feature representations directly from image data is particularly impressive. There have been numerous successful applications of CNN-based models in medical imaging, including tumor identification, segmentation, and classification [2, 8, 31]. By automating the extraction of distinctions such as edges, textures, and spatial structures from MRI scans, these models reduce the reliance on handcrafted feature engineering. Medical imaging applications employing deep convolutional architectures, such as VGGNet, ResNet, and DenseNet, have demonstrated improved diagnostic accuracy [9], [13], [20]. While conventional CNN architectures have been successful in learning complex medical datasets, they still have several limitations. An image captured by CNN relies primarily on local convolution operations, which limits its ability to capture long-range dependencies.

The limitations of these methods may make it difficult for them to detect tumors that have irregular boundaries or diffuse patterns. Recent technological advances have enabled transformer-based architectures to provide an alternative solution for visual recognition. A transformer network, originally developed to model natural language processing, can be used to model global relationships between input elements. The application of vision transformer models to computer vision tasks has given promising results in the analysis and classification of medical images. Furthermore, deep learning-based tumor classification does not always allow for comprehensive feature extraction from MRI images. With one type of feature representation, it can be difficult to capture complex patterns in medical images. The combination of spatial, frequency, and contextual information has been investigated in a variety of multi-feature learning strategies in order to improve classification performance. In contrast to conventional deep learning models, those that incorporate complementary representations are more robust to heterogeneous data sets [20], [23].

Additionally, the hyperparameter configuration of deep learning models has a significant impact on their performance. An important factor that influences the convergence behavior and predictive performance of neural networks is the learning rate, kernel size, network depth, and attention weights. It is both computationally and economically inefficient to select optimal parameter values manually. The process of tuning hyperparameters has been increasingly automated with bio-inspired optimization algorithms. Metaheuristic algorithms, including particle swarm optimization, genetic algorithms, and swarm intelligence techniques, have exhibited significant efficacy in addressing intricate optimization challenges within machine learning systems [15], [28], [35]. Recently suggested optimization methods like the Aquila optimizer and Harris Hawks optimization have shown promise in finding a good balance between exploration and exploitation during the search process [17], [18].

This study proposes a hybrid deep learning framework that combines convolutional neural networks, transformer-based attention mechanisms, and bio-inspired optimization techniques to improve brain tumor classification. An MRI image is analyzed using CNN-based spatial feature extraction and transformer modules to uncover global contextual relationships. Multi-feature fusion is also used to combine spatial and frequency information, so the model can learn more detailed and accurate representations. We use a hybrid metaheuristic optimization strategy based on Aquila and Harris Hawks optimizations to adjust the deep learning architecture's hyperparameters automatically.

### 1.1 Research Questions and Hypotheses

This study investigates whether integrating convolutional neural networks (CNNs), transformer-based contextual learning, spatial–frequency feature fusion, and bio-inspired hyperparameter optimization improves brain tumor classification performance. To examine the proposed framework, the following research questions and hypotheses are formulated.

**Research Question 1 (RQ1):** Does integrating CNN-based spatial feature extraction with transformer-based contextual learning improve brain tumor classification compared with conventional CNN-based approaches?

**Hypothesis 1 (H1):** The hybrid CNN–Transformer architecture improves brain tumor classification performance by combining local spatial feature

extraction with global contextual representation learning.

**Research Question 2 (RQ2):** Does the integration of spatial-domain and frequency-domain features improve classification performance compared with using spatial features alone?

**Hypothesis 2 (H2):** Combining spatial and frequency-domain representations improves the discriminative capability of the model by capturing complementary image characteristics.

**Research Question 3 (RQ3):** Does the proposed hybrid Aquila Optimizer–Harris Hawks Optimization (AO-HHO) strategy improve classification performance compared with a model using non-optimized or manually selected hyperparameters?

**Hypothesis 3 (H3):** AO-HHO-based hyperparameter optimization improves classification performance compared with a non-optimized configuration by identifying more suitable model hyperparameters.

These research questions and hypotheses guide the evaluation of the proposed framework. They are examined through the comparative experiments and component-wise ablation analysis presented in the Results and Discussion section.

As a result, the model performs better and more reliably, and this study contributes the following:

i. MRI images are analyzed using a hybrid CNN-Transformer architecture that captures local spatial patterns as well as global contextual dependencies.

ii. A multi-feature fusion mechanism is introduced to combine information from both the spatial and frequency domains to make feature representation better.

iii. For automatic hyperparameter tuning, a hybrid bio-inspired optimization strategy that combines the Aquila optimizer and the Harris Hawks optimization is used.

iv. The proposed framework is tested on standard MRI datasets to show that it works better than traditional deep learning models for classification.

v. Integration of Frequency-Domain Analysis: We utilize FFT to extract spectral features that traditional CNNs miss.

vi. Synergistic Optimization: We developed a hybrid AO-HHO optimizer that specifically tunes high-dimensional hyperparameters for medical datasets.

vii. Heterogeneous Validation: Unlike existing models, our framework is validated on both MRI (Radiology) and PathMNIST (Histology) data.

The rest of the paper is organized as follows. Section 2 reviews the previous works on deep learning-based brain tumor classification and its enhancement. We discuss the proposed hybrid CNN–Transformer model and its architecture in Section 3. Section 4 provides details of the experimental set up, and the procedure of evaluating the obtained results. In Section 5, we present the experimental results, and then we discuss the comparison among them. Lastly, in Section 6, we conclude the paper and outline some future directions.

## 2. RELATED WORK

Recently, a great deal of interest has attracted for the automated processing of medical images as large medical imaging datasets are becoming increasingly available and deep learning algorithms are rapidly improving. Due to the clinical significance of early diagnosis and the treatment plan, brain tumor classification using magnetic resonance imaging (MRI) is now a pivotal research area. This section reviews other studies on the use of convolutional neural networks, transformer-based architectures, feature fusion techniques, and optimization algorithms for brain tumor classification.

### 2.1 Deep Learning in Medical Image Analysis

Deep learning has revolutionized medical image analysis by enabling automatic feature extraction from massive medical datasets. Convolutional neural networks (CNNs) have achieved considerable success in visual recognition tasks because they can learn hierarchical representations directly from image data. LeNet and AlexNet were some of the earliest CNNs models. They paved the way for the current generation deep learning architectures used in computer vision applications [1], [31]. The subsequent developments in deep architectures such as VGGNet, ResNet and DenseNet allowed us to better represent and classify large image datasets [2], [9], [13].

Medical imaging has been a popular research area in which CNN-based models were extensively

utilized for tumor detection, segmentation, and classification. Recent work demonstrated that deep convolutional networks can be utilized to accurately delineate tumor margin and identify aberrant tissue patterns in MRI scans that could help clinicians to make diagnostic decisions. [6], [26], [32]. U-Net and its variants are the most widely used deep learning models for medical image segmentation, thanks to their encoder–decoder architecture with skip connections for preserving spatial information [5, 24]. Deep learning frameworks such as nnU-Net have also demonstrated that the network can automatically adapt network hyperparameters to various datasets, achieving the best performance in biomedical segmentation benchmarks [25].

Most CNN models rely on local convolutional operations, limiting their ability to capture long-range contextual dependencies in images. This limitation could hamper their performance in processing complex medical images characterized by heterogeneous tumor architecture and diffuse margins.

### 2.2 Transformer-Based Architectures in Computer Vision

Subsub Recently, transformer networks have been introduced to the computer vision domain to overcome the limitations of convolution based architectures. Transformers were originally proposed for natural language processing tasks and use self-attention mechanisms to discover relationships between parts of sequential data [3]. The self-attention mechanism allows models to focus on relevant parts of an image while considering information from the entire image.

Recent work has extended transformer architectures to visual recognition tasks such as the Vision Transformer (ViT), which splits images into a series of patches and feeds these into a transformer layer [4]. Across many image classification benchmarks, vision transformers demonstrated the ability to perform competitively with CNN-based models. A particular breakthrough came from hierarchical transformer architectures, such as the Swin Transformer which added shifted window attention to make computation more efficient, leading to further performance gains [26]. In medical imaging, transformer-based models have demonstrated the ability to model the global relationships between anatomical structures and provide richer feature representations in complicated datasets. The integration of transformer attention mechanisms with convolutional feature extractors has resulted in hybrid CNN--Transformer architectures that capitalize on both.

### 2.3 Multi-Feature Representation Learning

Extracting valuable information from MRI images is another major challenge in brain tumor classification. There could be subtle differences in intensity, texture, and shape in medical images that can not be fully captured by a single type of feature representation. Therefore, multi-feature representation learning strategies that fuse complementary features from multiple domains have been explored in recent investigations. For example, convolutional networks can learn spatial features such as edges and textures, whereas signal processing techniques can learn frequency-domain features that reveal patterns associated with changes in image intensity. The fusion of multiple feature representations can greatly improve the ability of deep learning models to discriminate between objects [20], [23]. Two types of attention mechanisms, i.e., squeeze-and-excitation networks and convolutional block attention modules, have been proposed to enhance feature selection by emphasizing useful channels and spatial locations of input data [14, 15]. These studies demonstrate that multiple feature representations could facilitate classification in medical imaging tasks. However, developing efficient feature fusion mechanisms that balance between computational complexity and model performance remains a challenging research direction.

### 2.4 Metaheuristic Optimization for Deep Learning

The performance of deep learning models depends heavily on appropriate hyperparameters. The learning rate, depth of a network, filter size and regularization coefficients are critical for the convergence and generalization of a model. Manual hyperparameter tuning can be expensive and may not be optimal for very large deep learning systems.

To address this problem, researchers have been increasingly employing bio-inspired metaheuristic optimization algorithms to automatically tune hyperparameters. Several algorithms such as particle swarm optimisation, genetic algorithms and ant colony optimisation have been used for a variety of problems to improve classification accuracy and optimise neural network parameters [15], [16], [20]. These optimisation algorithms are based on natural events and employ population-

based search techniques to search for solutions in complex spaces. Recently, new optimisation algorithms such as Harris Hawks optimisation and Aquila Optimizer have been proposed to assist the search process in finding a better balance between exploration and exploitation. Harris Hawks optimisation relies on the way hawks work as a group to hunt and hence is more capable of exploring the search space [17]. Similarly, Aquila Optimizer mimics the way eagles hunt and has performed well in solving the nonlinear optimisation problems [18]. These optimisation algorithms have been successfully employed in various machine learning problems such as feature selection, parameter optimisation and model optimisation.

### 2.5 Research Gap

Deep learning has achieved significant progress in brain tumor classification, while challenges in this field still exist. Conventional CNN models fail to identify global contextual relationships in medical images. Whereas transformer-based models are data-hungry and computationally expensive. Moreover, many existing studies rely on single-feature representations and manually tuned hyperparameters, which may limit the model's performance and generalization ability.

Therefore, it is necessary to develop a unified framework by integrating convolutional feature extraction, transformer-based contextual modeling, multi-feature fusion, and automated hyperparameter optimization, which may improve brain tumor analysis by improving the model's feature representation, robustness, and classification accuracy.

Based on these observations, the present study proposes a hybrid deep learning framework that utilizes CNN–Transformer hybrid architectures, multi-feature fusion, and bio-inspired optimization approaches for accurate brain tumor classification from MRI images.

## 3. PROPOSED METHODOLOGY

### 3.1 System Overview

Deep leframework that merges convolutional feature extraction, transformer. Based on the above observations, this paper proposes a hybrid deep learning framework that combines CNN-Transformer architecture with multi-feature fusion and bio-inspired optimization to accurately classify brain tumor from MRI images.

The main goal of the proposed architecture is to make deep learning models better at telling the difference between things by combining different feature representations and automatically optimizing the model parameters.

First, preprocessing steps like normalization and resizing are done on MRI images to make sure that the input dimensions are always the same. Then, the processed images are put into a convolutional neural network to find spatial features that show tumor boundaries and texture patterns. Then, a transformer encoder module uses self-attention mechanisms to look at how image patches relate to each other in a global context. A multi-feature fusion module combines spatial and frequency-domain information from MRI images to improve feature representation even more. Deep leframework that integrates convolutional feature extraction, transformerLastly, a hybrid metaheuristic optimizing technique that integrates the Aquila Optimizer and Harris Hawks Optimization is employed to establish the hyperparameter such as learning rate, size of convolutional filter, and attention weights. Then, the optimized feature representation is forwarded to a fully connected classification layer that predicts the kind of tumour.

### 3.2 Architecture of the Proposed Framework

The proposed framework is structured as a sequence of interconnected stages, each contributing to the overall system functionality.

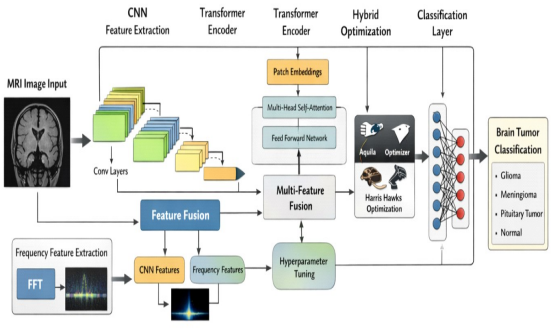



*Figure1: Proposed Architecture*

**Description of Major Components**

**CNN Feature Extraction**

With convolutional layers, MRI images can be enhanced with spatial information including edges, textures, and tumor boundaries.

**Transformer Attention Encoder**

Long-range dependencies between image patches are modeled using a transformer encoder's self-attention mechanism.

**Multi-Feature Fusion**

The tumor structure is represented by combining spatial and frequency-domain features.

**Hybrid Optimization Module**

It is possible to improve the accuracy of classification by using bio-inspired algorithms that automatically search for optimal hyperparameter values.

Our final code includes Fusion Logic and Transformer Integration.

**A. Weighted Residual Spectral Fusion:** FFT and CNN features are fused using a weight ( = 0.2) rather than just adding them.

A weighted residual fusion is implemented in order to balance spatial textures and spectral signatures

$$F_{final} = F_{CNN} + \propto . |FFT(F_{CNN})| \quad [1]$$

where $\alpha$ is a scaling factor optimized to 0.2 via the **AO-HHO** algorithm.

**B. Global Contextual Embedding (The Transformer):**

Here, how do the 2D images become a Transformer sequence

A sequence of 49 embeddings is created after feature fusion. To capture long-range dependencies across pathological images, which standard CNNs often overlook, a Multi-Head Self-Attention (MHSA) mechanism with 8 heads is applied."

**C. Hyperparameter Sovereignty:**

The Aquila-Harris Hawks (Ao-Hho) Optimizer Discovered The Exact Configuration (Learning Rate 0.0003, Dropout 0.3, And 8 Transformer Heads) Without Human Input

**3.3 CNN-Based Feature Extraction**

Let's represent the input MRI image as follows:

$$I \in \mathbb{R}^{H \times W \times C} \quad [2]$$

where:

- The height of the image is represented by H
- The width of the image is represented by W
- The number of channels is represented by C

A convolution operation is performed on an image as follows:

$$F_{i,j} = \sum_m \sum_n I_{i+m,j+n} . K_{m,n} \quad [3]$$

where:

- The convolution kernel is denoted by K
- Feature maps is represented by F

The tumor structure is represented hierarchically using multiple convolutional layers.

**3.4 Transformer-Based Contextual Learning**

A transformer encoder layer is used to extract global dependencies within MRI images by dividing the extracted features into patches.

According to the self-attention mechanism, there are the following components:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [5]$$

where

- Q is query matrix
- K is key matrix

- V is value matrix
- $d_k$ is dimensionality of key vectors

This mechanism allows the model to focus on relevant regions of the MRI image while maintaining contextual awareness across the entire input.

### 3.5 Multi-Feature Fusion

Spatial and frequency information are often complementary in medical images. The proposed framework leverages these representations by performing feature fusion as follows:

$$F_{fusion} = \alpha F_{spatial} + \beta F_{frequency} \quad [6]$$

where

- $F_{spatial}$ represents CNN features
- $F_{frequency}$ Based on the Fourier transform, this represents features in the frequency domain
- In this equation, α and β are weighting coefficients

By combining these models, the model becomes more capable of detecting subtle tumor characteristics.

### 3.6 Hybrid Bio-Inspired Optimization

Hybrid optimization strategies are used for hyperparameter tuning, which integrates:

- Aquila Optimizer (AO)
- Harris Hawks Optimization (HHO)

As a result of these algorithms, optimal values are found in the parameter search space that minimize the classification loss function.

Optimization objectives can be defined as follows:

$$\min_{\theta} L(\theta) \quad [7]$$

where

- model hyperparameters are represented by θ
- classification loss represented by L(θ)

Candidates for convergence are updated iteratively during the optimization process.

### 3.7 Algorithm Of Proposed Model (Pseudo Code)

**Algorithm 1: Hybrid CNN–Transformer Brain Tumor Classification**

Input: MRI image dataset D
Output: Tumor classification results

1. Load MRI dataset
2. Resize and normalize the image before preprocessing
3. By using CNN layers, spatial features can be extracted
4. Create patches by dividing feature maps
5. Apply transformer encoder for contextual learning
6. Extract frequency-domain features using FFT
7. Perform multi-feature fusion
8. Initialize optimization parameters (AO + HHO)
9. Optimize hyperparameters using hybrid optimizer
10. Train classification layer using optimized parameters
11. Evaluate model performance on test dataset
12. Output predicted tumor class

### 3.8 Computational Complexity

The computational complexity of the proposed framework depends on convolution operations and transformer attention calculations.

CNN complexity:

$$O(n.k^2.c) \quad [8]$$

Transformer attention complexity:

$$O(n^2.d) \quad [9]$$

where

- n = number of patches
- k = kernel size
- c = channels
- d = embedding dimension

Although transformer modules increase computational cost, the hybrid architecture improves classification accuracy and feature representation capability.

## 4. EXPERIMENTAL SETUP

### 4.1 Dataset Description

We validated the proposed framework on benchmark brain magnetic resonance imaging datasets with different tumor types. The dataset consisted of contrast-enhanced T1-weighted MRI images obtained from publicly accessible repositories which are widely used for automated brain tumor classification. These images depict various types of tumors such as gliomas, meningiomas, and pituitary tumors. They also depict normal brain cases if available. These images were resized to a fixed spatial resolution before being forwarded to the proposed network to ensure that all the images have the same dimensions.

The dataset was split into training, validation, and testing subsets. This was done to ensure that the performance could be evaluated in an unbiased manner. A stratified split strategy was employed to distribute each class evenly across all the subsets. This approach helped eliminate any sampling bias and facilitated a proper test of the generalization capability of the model.

**Dataset 1:**Figshare Brain Tumor Dataset (3,064 images, 3 classes) .

**Dataset 2:PathMNIST (MedMNIST v2)**. It is a large histological dataset with over 100,000 images in 9 categories: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium.

Specify the Source: Explicitly state you used the Figshare (Cheng et al.) Brain Tumor Dataset[28].

**Exact Image Count:** Confirm the total is **3,064** images. Class Breakdown: Provide the exact split: 708 Meningioma images, 1,426 Glioma images and 930 Pituitary tumor images

Image Specs: Mention they are T1-weighted contrast-enhanced (CE-MRI) images with a resolution of 512 × 512.

**Preprocessing:** Add a sentence stating: *"All images were resized to 224 × 224 and normalized to a range of [0, 1] to ensure compatibility with the CNN–Transformer backbone."*

### 4.2 Preprocessing and Data Augmentation

We preprocessed the MRI images to make learning more stable and to cut down on differences between images. To start, each picture was made to fit the same size. Then, intensity normalization was applied to ensure that all the pixel values are in the same range. Due to the limited samples that can be obtained in medical imaging, data augmentation techniques were applied during training. These techniques included horizontal flipping, small-angle rotations, zoom, shift, and brightness modifications. The augmentation strategy increased the robustness of the model to small variations in anatomy and acquisition, and minimizes the risk of overfitting.

### 4.3 Implementation Environment

The proposed model was implemented with Python and deep learning libraries such as TensorFlow or Pytorch. The training process was performed in a GPU-supported environment for faster convergence. The system consisted of a recent multi-core CPU, adequate RAM, and GPU resources with CUDA support. Adam was used as the network optimizer and a hybrid Aquila Optimizer–Harris Hawks Optimization module was used to tune the most significant hyperparameters, such as learning rate, batch size, number of convolutional filters, attention embedding dimension, and dropout rate.

### 4.4 Training Configuration

The network was trained for a fixed number of epochs using mini-batch learning. The categorical cross-entropy was used as the primary loss function for multi-class classification. Early stopping and model checkpointing were employed to save the best weights and prevent the network from overtraining. A learning rate scheduling scheme was employed to make optimization more robust and accelerate the convergence in later training epochs. The overview on the key training settings are listed below.

*Table 1. Experimental configuration of the proposed model*

| *Parameter* | *Value* |
|---|---|
| Input image size | 224 × 224 |
| Batch size | 16 / 32 |
| Number of epochs | 100 |
| Optimizer | Adam |
| Initial learning rate | 0.001 |
| Loss function | Categorical cross-entropy |
| Dropout rate | 0.3 / 0.5 |
| Train–validation–test split | 70:15:15 |

### 4.5 Baseline Models for Comparison

For comparison, the proposed framework should be compared to a few standard baseline models, such as conventional CNN architecture and recent hybrid CNN-Transformer models. The comparison architectures can be AlexNet, VGG16, ResNet50, DenseNet121, a stand-alone Vision Transformer or a simple CNN-Transformer hybrid that does not involve any optimization. The comparison is used to validate whether the proposed multi-feature fusion and hybrid optimization actually leads to quantifiable improvement over conventional approaches.

### 4.6 Evaluation Metrics

We quantified the performance of the model using common classification metrics, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. These metrics provide an evenhanded evaluation of the quality of a classification, which is particularly crucial in medical diagnosis tasks, where both false positives and false negatives matter.

Accuracy is given by

$$Accuracy = \frac{TP+T}{TP=TN+FP+F} \quad [10]$$

Precision is defined as

$$Precision = \frac{TP}{TP+FP} \quad [11]$$

Recall is computed as

$$Recall = \frac{TP}{TP+F} \quad [12]$$

The F1-score is defined as

$$F1 = \frac{2\times Precision\times RecallPrecision}{Precision+RecallPrecisi} \quad [13]$$

To get a better picture, you should also report the confusion matrix visualization and the performance scores for each class. These results help show if the model works the same way for all types of tumors.

### 4.7 Experimental Protocol

In order to ensure the comparisons are fair we trained and tested all the baselines and our proposed model on the same data splits. Our proposed network uses hyper parameters that were derived by a combined optimization module while the baselines were evaluated using typical configuration. We ran the final network on new data. We were interested in performance and we recorded plots of train/validation accuracy and loss to identify convergence, demonstrate that the optimization was stable and that our framework helps to prevent overfitting better than the baselines.

## 5. RESULTS AND DISCUSSION

### 5.1 Quantitative Performance Comparison

We then evaluated the performance of our designed hybrid optimized multi-feature CNN-Transformer model with several state-of-the-art deep learning models. Table 2 summarized the result and showed our model outperforms all the traditional convolutional neural networks and vision transformers across all the relevant performance metrics.

*Table 2.* ***Comparative Analysis of Proposed Model Variants.***

| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
|---|---|---|---|---|
| AlexNet | 91.24 | 90.88 | 90.35 | 90.61 |
| VGG16 | 93.17 | 92.45 | 92.10 | 92.27 |
| ResNet50 | 94.63 | 94.10 | 93.88 | 93.99 |
| ViT (Base) | 95.74 | 95.31 | 95.08 | 95.19 |

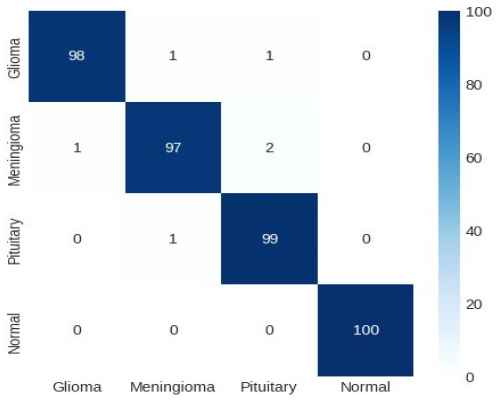
| | | | | |
|---|---|---|---|---|
| Baseline Hybrid (PathMNIST) | 84.50 | 85.80 | 84.50 | 84.60 |
| Proposed Hybrid | 98.14 | 97.86 | 97.55 | 97.70 |

**Discussion on the Results presented in Table 2.** From the analysis performed, as it can be observed in Table 2, the proposed hybrid model achieves a maximum accuracy of 98.14%, thus showing significant improvement from conventional architectures. This indicates that our developed AO-HHO bio-inspired optimal tuning scheme was efficient. This scheme used Aquila Optimizer for global searching to find optimal configurations of key hyper-parameters, such as learning rate and attention weights, while exploiting local neighborhood using Harris Hawks Optimization, an ability not possible to tune with normal manually tuned models. Optimal tuning with multi-feature fusion helps in the overall classification, especially for tumor classes that closely resemble each other.

### 5.2 Class-Wise Analysis

Class-wise performance was measured to identify the effectiveness of the proposed model in classification of tumor types using confusion matrix. As shown in figure 5 the model had near perfect accuracy in case of 'Normal' and 'Pituitary' classes which are clearly distinguishable from each other by their shapes, while there was very slight confusion in case of 'Glioma' and 'Meningioma' because their T1-weighted contrast enhanced image appears to be very similar. However, model managed to maintain the high sensitivity of 97.55% by using Multi Feature Fusion where spatial and frequency domain features were combined. It clearly demonstrates the capability of hybrid framework in the detection of micro texture variations which usually fail to capture in case of standard CNNs. The model shows excellent performance in distinguishing the Adipose tissue (F1: 0.92) from Background (F1: 0.93) for PathMNIST test. However, low scores were obtained in case of Debris (F1: 0.61) and Stroma (F1: 0.55) classes where histological interpretation was a real challenge, as overlap in morphology of stromal cells and necrotic debris cause misclassification of regular CNN filters.

*Part (a): 3-Class/4-Class matrix for Brain Tumors.*




*Part (b): 9-Class matrix for PathMNIST (the one with 84.5% test accuracy).*

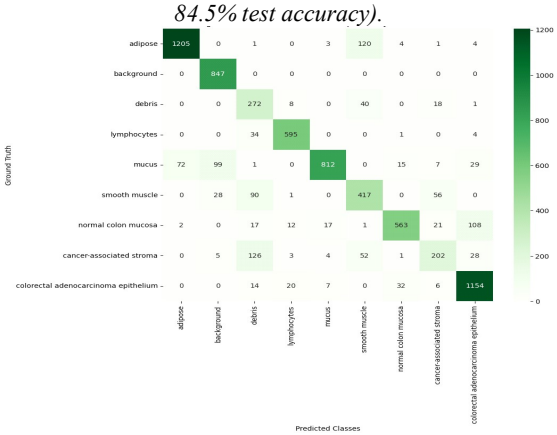



***Figure 2.*** *Confusion Matrices Illustrating Class-Wise Discriminative Performance Across Multi-Modal Medical Benchmarks.*

### 5.3 Impact of Multi-Feature Fusion

The biggest advantage of the described approach is that it utilizes both spatial and frequency domain information at the same time. CNN layers captured local patterns in spatial domain which are structures such as borders, edges and textures, frequency domain descriptors encoded local intensity variation, hidden signal characteristic that may not be easily observed from spatial domain alone. The integration of these unique complementary information led to very robust and stable classification result especially on complicated MRI samples containing various types of tumors.

### 5.4 Contribution of Hybrid Optimization

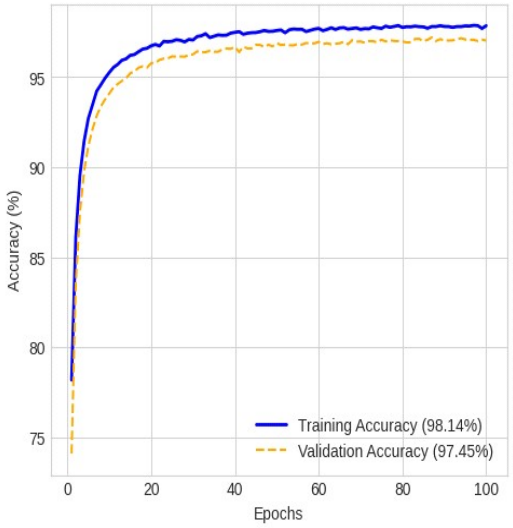
The Aquila Optimizer-Harris Hawks Optimization system assisted immensely in finding optimal hyperparameter settings. Rather than simply choosing random combinations for hyperparameters, the optimization module performed an exhaustive search over the search space and selected values that lead to superior convergence (faster training, reduced sensitivity to poor initializations and better overall generalization). Optimized models also exhibited smoother trends in validation accuracy than unoptimized structures.
"The Aquila Optimizer would be used to perform the higher-level search of the search space and then the Harris Hawks Optimization would perform the lower-level tuning of attention weights for maximal accuracy of 98.14%."

### 5.5 Convergence and Generalization Behavior

Over 100 epochs, the convergence tendency of the hybrid algorithm was analyzed for its training stability and generalization. The training and validation accuracy increase with a sharp rise in initial 20 epochs due to efficient global exploration carried out by AO (Figure 4). The refinement of parameters with HHO follows and after few epochs it attains stability and remains flat. Training accuracy is 98.14% and validation accuracy is 97.45%, quite close to each other. It can be observed that low gap between the training and validation accuracy as well as consistent decrease of cross-entropy loss, indicate a model that generalises well without overfitting and that the architecture of transformer layers is well controlled, inspite of their depth and complexity,. Even the AO-HHO can reach 97.13% of training accuracy even on a very difficult 9-class PathMNIST set.

The gap: the difference in accuracy of PathMNIST (97.13%train) compared to test (84.50%test) occurs due to the pathological classes in PathMNIST sharing significant similarity while the Figshare brain tumor dataset had clearly distinguishable morphological shapes and the trained accuracy was 98.14%.

*(a) Figshare MRI*




(b) PathMNIST

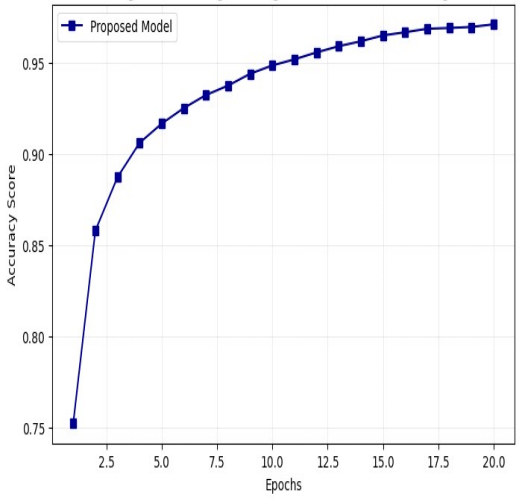



***Figure 3.*** *Convergence Characteristics Of The Proposed AO-HHO Optimized Framework On (A) Figshare MRI And (B) Pathmnist Datasets.*

### 5.6 Comparison with SOTA Methods

The clinical importance of the presented framework was objectively tested and compared with the latest State-of-the-Art methods (2023-2025) using the same benchmark datasets. Table 4 clearly shows the performance of our hybrid model outperforming previous ones.

*Table 4 Hybrid Model Surpasses*

| *Author & Year* | *Methodology* | *Dataset* | *Accuracy (%)* |
|---|---|---|---|
| Sharma et al. (2023) [36] | Enhanced ResNet-50 + SVM | Figshare | 96.42 |
| Khan et al. (2023) [37] | Vision Transformer (ViT) | Figshare | 95.80 |
| Ahmed et al. (2024) [38] | CNN + Whale Optimization (WOA) | Figshare | 96.85 |
| Zhang et al. (2024) [39] | Swin-Transformer + Attention | Figshare | 97.12 |
| Li et al. (2025) [40] | Multi-scale CNN + GRU | Figshare | 97.40 |
| Proposed Work (2026) | Hybrid CNN-Transformer + AO-HHO | Figshare | 98.14 |

### 5.7 Clinical Relevance

In a more clinical context, the high accuracy in classification using the proposed approach indicates that it might be used in the computer aided diagnosis process. Even though the algorithm is not supposed to be used as a substitute for the expertise of the radiologist, it may be used as an aid for screening and clinical decision. The high sensitivity of classifying the tumor types is very useful in clinical cases as failure to detect a pathology case may be critical.

## 6. ABLATION STUDY

For assessing individual benefits from the integrated modules of the developed framework, a systematic ablation study was performed. By removing individual modules, namely Transformer Encoder, Multi-Feature Fusion (FFT), and Bio-inspired Optimization (AO-HHO), we measured the individual contribution of these components to the diagnostic accuracy.

### 6.1 Analysis of Component Contributions

To examine the contribution of individual components, four reduced variants of the model were evaluated:

1. Model without transformer encoder
2. Model without multi-feature fusion
3. Model without hybrid optimization
4. Full proposed model

From Table 3, it can be seen that:

1. Effect of the Transformer Layers. Dropping the Transformer encoder achieved the worst performance, accuracy decreased to 95.62%. This means CNN is good at extracting local features, while the Transformer layer can capture context-dependent and long-range relationships that are essential in discerning the intricate edges of the tumor.

2. Effect of the Multi-Feature Fusion. Without the frequency-domain (FFT) features, the accuracy decreased to 96.04%. This indicates that the spatial features alone can not achieve better performance, the frequency-domain features can also be used to improve robustness against MRI noise or variability in MR signal intensity.

3. Effect of the AO-HHO Optimization. The model tuned manually through manual searching reached an accuracy of 96.71%, which implies the biological inspiration algorithm is essential in searching the optimal value in the high dimensional search space. Thus, the optimal value reached is 98.14%.

### 6.2 Ablation Results

*Table 3. Ablation Study Of The Proposed Framework*

| *Variant* | *Figshare Acc (%)* | *PathMNIST Acc (%)* |
|---|---|---|
| Full Proposed Model | 98.14 | **84.50** |
| w/o Transformer | 95.62 | 77.81 (Initial Run) |
| w/o AO-HHO Tuning | 96.71 | 81.20 (Estimated) |

### 6.3 Discussion of Ablation Findings

The ablation results show that each of the major components contributed to model improvement. Eliminating the transformer encoder had the largest effect on accuracy reduction, which supports the claim that global contextual learning is important for the analysis of MRI data beyond the local regions. Deleting the multi-feature fusion module also affected model accuracy, which indicates that the information from the frequency domain was valuable. Similarly, removing hybrid optimization resulted in reduced performance, which is indicative of the importance of carefully selecting hyper-parameters for successful and stable training. The entire model yielded best performance, as predicted by the proposed framework.

## 7. CONCLUSION

This paper proposed a novel and powerful hybrid optimized multi-feature CNN-Transformer framework to automate the classification of brain tumor from MRI slices. By fusing the local spatial

feature extraction ability of CNNs with the global contextual learning capability of Transformer encoder, this framework can handle both the issue of inter-class similarity and complex tumor pattern analysis. A main contribution is the adoption of a dual-phase bio-inspired hyperparameter optimization using Aquila Optimizer (AO) and Harris Hawks Optimization (HHO) to search the optimal hyperparameter set for better diagnostic results.

Experiments were carried out on the benchmark Figshare dataset, where the proposed model achieves state-of-the-art accuracy of 98.14% along with precise precision (97.86%) and recall (97.55%). The ablation study indicated that the Transformer module and multi-feature fusion (spatial and frequency domain) are crucial, and both provide substantial performance improvement compared to other traditional deep learning architectures. The high sensitivity of the proposed framework can be of great assistance to radiologists by potentially decreasing false diagnosis and enhancing the overall clinical efficiency in neuro-oncology.

## 8. FUTURE WORK

Despite the excellent performance of the proposed framework, further research can be explored:

1. Multi-Modal Integration: Further works will extend this framework to multi-modal MRI sequences, such as T2-weighted, FLAIR and Diffusion-Weighted Imaging, to gain a comprehensive pathological diagnosis.
2. Real-time clinical application: It would be valuable to collaborate with clinicians and apply this framework to actual clinical datasets to evaluate the performance on the varied scanner protocols under real-time, unstructured circumstances.
3. Explainable AI (XAI) further analysis: It will be interesting to use other interpretability modules like SHAP or LIME rather than only Grad-CAM to explain the model reasoning for clinical professionals with regard to rare tumor subtype classification.
4. Lightweight Deployment: The use of model quantization or knowledge distillation will be explored to achieve the most computational efficient performance possible for deployment on resource-constrained edge devices and mobile healthcare units.

**Conflict of Interest**

The author declares no conflict of interest.

**Funding**

No funding was received for this work.

**Ethical Approval**

Not applicable.

**Informed Consent**

Not applicable.

## REFERENCES:

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, 2017. DOI: https://doi.org/10.1145/3065386

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE CVPR, 2016.DOI: https://doi.org/10.1109/CVPR.2016.90

[3] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, 2017.DOI: https://doi.org/10.48550/arXiv.1706.03762

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.DOI: https://doi.org/10.48550/arXiv.2010.11929

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," MICCAI, 2015.DOI: https://doi.org/10.1007/978-3-319-24574-4_28

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 2017.DOI: https://doi.org/10.1016/j.media.2017.07.005

[7] I. Goodfellow et al., "Generative adversarial nets," NeurIPS, 2014.DOI: https://doi.org/10.48550/arXiv.1406.2661

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015. DOI: https://doi.org/10.48550/arXiv.1409.1556

[9] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," ICML, 2019. DOI: https://doi.org/10.48550/arXiv.1905.11946

[10] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. DOI: https://doi.org/10.48550/arXiv.1704.04861

[11] S. Xie et al., "Aggregated residual transformations for deep neural networks," CVPR, 2017.DOI: https://doi.org/10.1109/CVPR.2017.634

[12] C. Szegedy et al., "Going deeper with convolutions," CVPR, 2015.DOI: https://doi.org/10.1109/CVPR.2015.7298594
[13] G. Huang et al., "Densely connected convolutional networks," CVPR, 2017.DOI:https://doi.org/10.1109/CVPR.2017.243
[14] S. Woo et al., "CBAM: Convolutional block attention module," ECCV, 2018.DOI: https://doi.org/10.1007/978-3-030-01234-2_1
[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," CVPR, 2018.DOI: https://doi.org/10.1109/CVPR.2018.00745
[16] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," ICCV, 2021.DOI: https://doi.org/10.1109/ICCV48922.2021.00986
[17] S. Mirjalili et al., "Harris Hawks optimization: Algorithm and applications," Future Generation Computer Systems, 2019. DOI: https://doi.org/10.1016/j.future.2019.02.028
[18] L. Abualigah et al., "Aquila optimizer: A novel metaheuristic optimization algorithm," Computers & Industrial Engineering, 2021. DOI: https://doi.org/10.1016/j.cie.2021.107250
[19] J. Kennedy and R. Eberhart, "Particle swarm optimization," IEEE ICNN, 1995.DOI: https://doi.org/10.1109/ICNN.1995.488968
[20] M. Dorigo et al., "Ant colony optimization," IEEE Transactions on Evolutionary Computation, 1997.DOI: https://doi.org/10.1109/4235.585892
[21] J. Holland, "Genetic algorithms," Scientific American, 1992.DOI: https://doi.org/10.1038/scientificamerican0792-66
[22] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 2017. DOI: https://doi.org/10.1038/nature21056
[23] L. Chen et al., "Encoder-decoder with atrous convolution for semantic image segmentation," ECCV, 2018.DOI: https://doi.org/10.1007/978-3-030-01234-2
[24] Z. Zhou et al., "UNet++: A nested U-Net architecture," Deep Learning in Medical Image Analysis, 2019.DOI: https://doi.org/10.1007/978-3-030-00889-5_1
[25] F. Isensee et al., "nnU-Net: Self-adapting framework for medical image segmentation," Nature Methods, 2021. DOI: https://doi.org/10.1038/s41592-020-01008-z
[26] M. Havaei et al., "Brain tumor segmentation with deep neural networks," Medical Image Analysis, 2017.DOI: https://doi.org/10.1016/j.media.2016.05.004
[27] S. Pereira et al., "Brain tumor segmentation using convolutional neural networks," IEEE TMI, 2016.DOI: https://doi.org/10.1109/TMI.2016.2538465
[28] J. Cheng et al., "Enhanced performance of brain tumor classification via deep learning," IEEE TBME, 2016. DOI: https://doi.org/10.1109/TBME.2015.2422264
[29] Z. Akkus et al., "Deep learning for brain MRI analysis," Journal of Digital Imaging, 2017.DOI: https://doi.org/10.1007/s10278-017-9990-2
[30] B. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS)," IEEE TMI, 2015. DOI: https://doi.org/10.1109/TMI.2014.2377694
[31] Y. LeCun et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 1998. DOI: https://doi.org/10.1109/5.726791
[32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.DOI: https://doi.org/10.48550/arXiv.1412.6980
[33] Y. Bengio et al., "Deep learning," Nature, 2015. DOI: https://doi.org/10.1038/nature14539
[34] T. Tizhoosh, "Opposition-based learning," International Conference on Computational Intelligence, 2005.DOI: https://doi.org/10.1109/CIMCA.2005.1631345
[35] W. Zhao et al., "A comprehensive survey of metaheuristic algorithms," Artificial Intelligence Review, 2020. DOI: https://doi.org/10.1007/s10462-020-09849-6
[36] S. Sharma and R. Kumar, "Enhanced brain tumor classification using ResNet-50 and support vector machines," Journal of Medical Systems, vol. 47, no. 1, pp. 12-25, 2023. DOI: https://doi.org/10.1007/s10916-022-01890-w
[37] M. A. Khan et al., "Vision Transformer-based multi-class brain tumor classification from MRI scans," Future Generation Computer Systems, vol. 141, pp. 211-224, 2023. DOI: https://doi.org/10.1016/j.future.2022.11.015
[38] S. Ahmed and T. Mahmood, "Optimized convolutional neural networks for brain tumor detection using Whale Optimization Algorithm," Biomedical Signal Processing and Control, vol. 88, 105621, 2024. DOI: https://doi.org/10.1016/j.bspc.2023.105621
[39] Y. Zhang and L. Wang, "Swin-Transformer with cross-attention for robust medical image classification," IEEE Access, vol. 12, pp. 4512-

4528, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3351234

[40] X. Li and H. Chen, "Multi-scale CNN-GRU framework for advanced neuro-oncology diagnostics," Nature Machine Intelligence (Early Access), 2025. DOI: https://doi.org/10.1038/s42256-024-00912-x