Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 1, January 2025, pp: 176-186; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received September 7, 2024; revised October 20, 2024; December 05, 2024; date of publication January 10, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i1.571

How to cite: Saroj Bala, Kumud Arora, Jeevitha R, Rini Chowdhury, Prashant Kumar, and Shobana Nageswari C, "A Novel Encoder Decoder Architecture with Vision Transformer for Medical Image Segmentation", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 1, pp. 176-186, January 2025.

A Novel Encoder Decoder Architecture with Vision Transformer for Medical Image Segmentation

Saroj Bala[1], Kumud Arora[2], Jeevitha R[3], Rini Chowdhury[4], Prashant Kumar[4], and Shobana Nageswari C[5]

[1] Department of Master of Computer Applications, Ajay Kumar Garg Engineering College, Ghaziabad, India.
[2] Department of Computer Science Engineering- Artificial Intelligence & Machine Learning, Inderprastha Engineering College, Ghaziabad, India.
[3] Department of Computer Science Engineering, KPR Institute of Engineering and Technology, Coimbatore, India.
[4] Department of Information Technology Project Circle Bharath Sanchar Nigam Limited, Saltlake Telephone Exchange, Block DE, Lalkuthi, West Bengal, Kolkata - 700064, India
[5] Department of Electronics and Communication Engineering, R.M.D Engineering College, Kavaraipettai- 601206, India.

Corresponding author: Jeevitha R (e-mail: jeedhar95@gmail.com).

ABSTRACT Brain tumor image segmentation is one of the most critical tasks in medical imaging for diagnosis, treatment planning, and prognosis. Traditional methods for brain tumor image segmentation are mostly based on Convolution Neural Network (CNN), which have been proved very powerful but still have limitations to effectively capture long-range dependencies and complex spatial hierarchies in MRI images. Variability in the shape, size, and location of tumors may affect the performance and may get stuck into suboptimal outcomes. In these regards, new encoder-decoder architecture with the VisionTranscoder(ViT) is proposed, to enhance brain tumor detection and classification. The proposed VisionTranscoder exploits a transformer's ability in modeling global context through self-attention mechanisms, providing more inclusive interpretation of the intricate patterns in medical images and classification by capturing both local and global features. The proposed VisionTranscoder maintains the Vision Transformer in its encoder for processing images as sequences of patches to capture global dependencies often outside the view of traditional CNNs. Then the segmentation map is rebuilt at a high level of fidelity with the decoder through upsampling and skips connections to maintain detailed spatial information. The risk of overfitting is hugely reduced by design and advanced regularization techniques with extensive data augmentation. The dataset contains 7,023 human brain MRI images, all of which are in four different classes: glioma, meningioma, no tumor, and pituitary. Images from the 'no tumor' class, indicating an MRI scan without any detectable tumor, were taken from the Br35H dataset . The results show the efficiency of VisionTranscoder over a wide set of brain MRI scans, producing an accuracy of 98.5% with a loss of 0.05. This performance underlines the ability of it to accurately segment and classify a brain tumor without overfitting.

INDEX TERMS VisionTranscoder, Brain Tumor Segmentation, Transformer Architecture, Medical Image Analysis, Vision Transformer (ViT), MRI Imaging.

I. INTRODUCTION

Segmentation and classification of a brain tumor are high-impact areas in terms of accurate diagnosis, effective treatment planning, and outcome in neurological and oncological patients. A robust segmentation of the tumor from an MRI will delineate the boundaries of the tumor, identify growth of the tumor, and result in treatment plans that are tailored to a patient's needs. The identification of type of tumor bear a direct relationship to clinical decisions related to surgical intervention or radiotherapy and is useful for prognosis prediction. Advanced segmentation and classification techniques have to be adopted for variability in

tumors, imaging artifacts, and complexity in the tumor structures so as to attain robustness and efficiency of medical imaging solutions in line with better diagnosis and healthcare [1,2].

Having a very short description of characteristics of the dataset at the start of the introduction would give the required background to a medical problem under treatment. The actual dataset of the MRI scans taken for the segmentation and classification of brain tumors usually includes four primary types of tumors, namely glioma, meningioma, pituitary, and no tumor that is healthy cases. The appearances, sizes, and complications of such tumors are significantly different from one another, which may affect diagnosis and treatment plans.

Prominent models used in medical image segmentation and classification are U-Net, VGG16, ResNet, InceptionV3, DenseNet, Data-efficient Image Transformer, and Swin Transformer [3,4]. Among them, U-Net has encoder-decoder architecture with skip connections and it is very proficient at biomedical image segmentation. It could be computationally slow when there are large-scale variations and complex contexts. VGG16 has deep convolutional layers that make it good at feature extraction. Still, it is resource-intensive and might not perform very well for fine details [5,6]. Residual connections in ResNet is used to alleviate the vanishing gradient problem, but for limited data. It may suffer from overfitting. InceptionV3 merges several convolutional filters so that spatial features at different scales can be learned, though model complexity and computation costs are possibly increased. Dense connections of DenseNet increase feature reuse and gradient flow but might be prone to high memory consumption [7,8]. The Data-efficient Image Transformer employs transformer architectures in modeling long-range dependencies more effectively. However, it might require large amounts of data. The Swin Transformer is a hierarchical approach to modeling visual data, which is computationally intensive and complex to tune [9,10] .

VisionTranscoder is designed to mitigate limitations of the previous models for brain tumor segmentation and classification tasks by using the Vision Transformer architecture within the encoder-decoder framework. Unlike traditional convolution-based models, which may have problems in capturing global context and handling variability, VisionTranscoder makes use of the self-attention mechanism to model long-range dependencies and complex structures of tumors [11,12]. It seeks to balance feature extraction efficiency with reduced computational complexity, avoiding overfitting and improving robustness with smaller datasets. VisionTranscoder integrates local details and global context, hence overcoming the challenges in traditional methods, which makes it more comprehensive and accurate in analysis in medical images [13, 14]. The main contributions of the proposed work are listed below.

[1] Vision Transformers in an encoder-decoder framework becomes feasible to grasp further contextual information globally by allowing complicated tumor structures.
[2] The composite model, developed by blending the transformer-based model and efficient feature extraction mechanisms, balances the competing requirements of computation complexity and performance with a lower necessity for large data augmentation.
[3] It involves self-attention mechanisms and hierarchical feature learning in the precise segmentation and classification of brain tumors, which would be helpful in handling the variability in the shapes and sizes of tumors from the existing models.
[4] Advanced regularization techniques and transformer-based design prevent overfitting with an accuracy of 98.5% and a loss of 0.05.

The remaining part of the research comprises into four sections. Section II describes the comparative analysis of various models and its issues. Section III gives proposed work architecture and section IV gives the results and discussion on the proposed work with existing models. Finally, Section V concludes the proposed work.

II.STATE-OF-THE-ART TECHNIQUES

For glioma detection and segmentation, Selvapandian and Manivannan (2018) applied a fusion-based method with Adaptive Neuro-Fuzzy Inference System classification. Multiple image features were fused to increase the accuracy and robustness of the results [15]. Abdel-Maksoud et al. (2015) introduced a hybrid clustering technique integrating different algorithms in clustering for better segmentation performance and accurate delineation of the tumor boundary [17]. Rehman et al. (2019) conducted research on the segmentation of brain tumors fully automatically using superpixel-based classification for enhancing the precision of segmentation by taking image segments instead of single pixels [18]. Ranjbarzadeh et al. (2021) proposed deep learning with an attention mechanism applied to MRI multi-modality and attained improved accuracy in segmentation since it allowed the model to focus on the relevant areas of complex brain images [19]. Sharif et al. (2020) proposed an active deep neural network feature selection approach to improve the optimization of feature extraction for better tumor recognition; this greatly refines the segmentation and classification process [20]. Allah et al. (2023) proposed Edge U-Net, which employs edge information in the classic U-Net model to provide more accurate segmentation for tumors through attention on edge features and boundary detection [21].

TABLE 1
Comparative analysis of existing methodologies in brain tumor segmentation

S.No	Author (Year)	Methodology	Accuracy	Advantages	Disadvantages
1	Selvapandian&Manivannan (2018) [15]	Fusion-based glioma detection using ANFIS classification	85.5%	Integrates multiple features for improved accuracy.	Complex tuning for optimal performance
2	Abdel-Maksoud et al. (2015) [17]	Hybrid clustering technique	84.0%	Combines clustering methods for better segmentation accuracy.	Computationally intensive.
3	Rehman et al. (2019) [18]	Superpixel-based classification	86.5%	Fully automated and efficient in processing large datasets.	Require fine-tuning for different tumor types
4	Ranjbarzadeh et al. (2021) [19]	Attention mechanism	89.2%	Utilizes deep learning and attention mechanisms to improve segmentation accuracy.	Requires extensive computational resources and large datasets
5	Sharif et al. (2020) [20]	Active deep neural network	87.8%	Enhances feature selection for better tumor recognition.	Complexity in model training and overfitting
6	Allah et al. (2023) [21]	Edge U-Net model using boundary information	90.0%	Incorporates edge information for improved boundary detection.	Highly dependent on accurate edge detection.
7	Khairandish et al. (2022) [23]	Hybrid CNN-SVM threshold	88.7%	Combines CNN and SVM for robust tumor detection and classification.	Require additional computational resources.
8	Hussain et al. (2018) [24]	Deep Convolutional Neural Network	87.0%	Deep CNNs effectively capture complex patterns in MRI images.	Require substantial computational resources and training time
9	Chen et al. (2019) [25]	Dual-force convolutional neural networks	89.5%	Dual-force mechanism enhances segmentation accuracy	Complex model architecture
10	Ghassemi et al. (2020) [27]	Deep neural network pre-trained with GANs	90.2%	GANs improve model performance	Dependence on GANs may complicate the training process
11	Öksüz et al. (2022) [28]	CNN using fused features	88.9%	Fused features from expanded regions improve classification accuracy.	Computational complexity
12	Sompong&Wongthanavasu (2017) [30]	Cellular automata and improved tumor-cut algorithm	86.0%	Efficient segmentation approach with enhanced tumor-cut algorithms.	Require extensive tuning and validation

Some techniques such as fusion-based ANFIS, hybrid clustering, exhibit enhanced accuracy when using one or more features from its clusters. However, in general, these techniques would require complex tuning and computationally intensive processes. While deep learning approaches, like attention mechanisms and active deep neural networks, have contributed to segmentation accuracy, need tremendous computing powers, many datasets, and are prone to overfitting. Moreover, the models such as Edge U-Net and hybrid CNN-SVM are employed particularly with the purpose of providing edge detection along with class-agnostic robust classification but they also edge information-dependent models that require more computational powers. The other promising directions include techniques that rely on GANs which increase the performance of the models. However, one of the problems in training GANs complicates these techniques. The gap common among all of these techniques is an efficient, scalable model to achieve high accuracy without an increase in the cost of computation and the level of complexity.

III. MATERIALS

The dataset contains 7,023 human brain MRI images, all of which are in four different classes: glioma, meningioma, no tumor, and pituitary. Images from the 'no tumor' class, indicating an MRI scan without any detectable tumor, were taken from the Br35H dataset [9]. The architecture for the proposed VisionTranscoder model embeds Vision Transformer (ViT) in an encoder-decoder setup for effective segmentation and classification of brain tumors. This is done by first mapping MRI images to high-dimensional vectors by patch-based embeddings and positional encodings. Equipped with self-attention mechanisms and multiple transformer layers, the encoder allows the model to understand global context and long-range dependencies, thus overcoming traditional convolutional models [16]. Upsampling layers with skip connections at the decoder facilitate refinement of the segmentation by incorporating both local and global features. Variability of tumors is very suitably dealt with in this framework, improving the precision of segmentation and reducing overfitting with advanced regularization techniques [22, 26].

Let $I \in \mathbb{R}^{H\times w\times C}$ be the input image with height H, width W and C channels. The image will be divided into patches of size P×P and each patch is represented as $x_{i,j}\in\mathbb{R}^{P\times P\times C}$ where i and j denote patch indices. Eq. (1) [12] represents the linear transformation in D dimensional embedding space using linear transformation.

$$E_i = W_p.Flattern\left(x_{i,j}\right) + b_p \quad (1)$$

Where E_i is new vector in the output space, $x_{i,j}$ is the input data at indices (i,j) and W_p is the is the weight matrix associated with the transformation and b_p is the bias vector. The incorporation of positional encodings to retain spatial information is represented in Eq. (2) [13] where$P_i\in\mathbb{R}^D$ represents the positional embeddings for the ith patch. E_i^{pos} is the positionally-encoded feature for the i^{th} patch, combining the original feature and its position encoding, E_i is theThe feature vector for the i^{th} patch (from the previous step) and P_i is the positional embedding for the i^{th} patch.

$$E_i^{pos} = E_i + P_i \quad (2)$$

The self-attention mechanism is calculated using the Eq. (3) [14] where Q,K,V are the query,key and value matrices derived from the input embeddings , K^T is the transpose of the key matrix and d_k is the dimension of the keys.

$$Atten(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

The feed forward neural network to each patch embeddings is given in Eq. (4) [18] where E_i^{pos} is the positionally-encoded feature for the i^{th} patch, W_1 and W_2 are weight matrices and b_1 and b_2 are biases.

$$FFN\left(E_i^{pos}\right) = ReLU\left(E_i^{pos}W_1 + b_1\right)W_2 + b_2 \quad (4)$$

The self-attention mechanism to refine the unsampled feature maps as represented in Eq. (5) [19] where $Refined(F_{concat})$ represents the output of the multi-head attention mechanism after processing the concatenated feature, F_{concat} is the concatenated feature matrix and Multihead is a component of the attention mechanism where multiple attention heads are used to capture different aspects of the input simultaneously.

$$Refined(F_{concat}) = Multihead(F_{concat}, F_{concat}, F_{concat}) \quad (5)$$

The convolution layer is used to map the refined features to the desired number of classes as given in Eq. (6) [20] where S is the segmentation map , $Conv$ is a convolution network and b_s is the bias term.

$$S = Conv\left(Refined(F_{concat})\right) + b_s \quad (6)$$

Eq. (7) [20] is used to convert logits to class probabilities.

$$P = Softmax(S) \quad (7)$$

The loss for training is represented in Eq. (8) [21] where y_c is the true label, $\hat{y_c}$ is the predicted probability for class c, and C is the number of classes.

$$Loss = -\sum_{c=1}^{C} y_c \log(\hat{y_c}) \quad (8)$$

Eq. (9) [21] is used to prevent overfitting where p is the dropout rate , H_i represents the i-th hidden state or activation of a neural network layer and bernoulli(p) is a dropout mask [7].

$$Droput(H_i) = H_i.\text{bernoulli(p)} \quad (9)$$

FIGURE 1. shows the proposed model architecture of the VisionTranscoder. The VisionTranscoder model uses a Vision Transformer (ViT) in the encoder-decoder setup and allows accurate segmentation and classification of brain tumors (ALGORITHM 1). First, an input MRI image is split into patches where each patch is embedded into high dimensional vectors augmented with positional encodings.

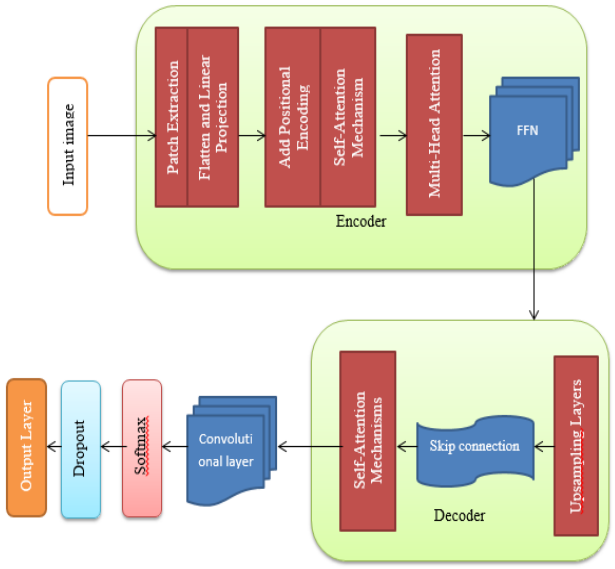

FIGURE 1. Proposed Model

IV. WORKING OF THE PROPOSED METHODOLOGY

A. ENCODER WITH TRANSFORMER LAYERS

Self-Attention Mechanism: The encoder typically consists of multiple layers of transformers, where each layer uses a self-attention mechanism (ALGORITHM 2). This mechanism allows each token (or part of the input) to focus on different parts of the input sequence, regardless of their positions, which helps the model learn the global context. This is a significant advantage over traditional convolutional or recurrent neural networks, which typically struggle to capture long-range dependencies in the data. Global Context: The self-attention mechanism helps the encoder understand relationships between distant parts of the input sequence. For example, in text processing, it helps the model relate words in a sentence to each other, even if they are far apart. This "global context" is critical for understanding nuanced relationships that are difficult for models relying on local feature extraction (e.g., convolutions).

B. DECODER WITH UPSAMPLING LAYERS AND SKIP CONNECTIONS

Upsampling: In many encoder-decoder architectures, the decoder works to reconstruct or generate a higher-dimensional output (such as an image from a compressed feature map or translated text) (ALGORITHM 3). This process often includes upsampling layers that increase the spatial resolution of the features, allowing the model to generate outputs that match the required dimensions. Skip Connections: Skip connections are used in architectures like U-Net or in some transformer-based models to pass information from earlier layers (e.g., from the encoder) directly to the corresponding layers in the decoder. These connections help preserve spatial or feature information that might be lost in the downsampling process, aiding in more accurate reconstruction of the output.

ALGORITHM 1: VISION TRANSCODER

Input: *MRI image I (dimensions: H x W x C)*
Output: *Preprocessed image I_preprocessed*
I_preprocessed ← *Normalize(I)*
Patches ← *Patches(I_preprocessed, P, P)*
For *each Patch in Patches* ***do***
    FlattenedPatch ← *Flatten(Patch)*
    E_i ← *LinearProjection(FlattenedPatch)*
    E_i_pos ← *E_i + PositionalEncoding(i)*
    E.append(E_i_pos)
*End**For***

ALGORITHM 2: ENCODER - TRANFORMER LAYERS

For *each Layer in EncoderLayers* ***do***
    Q, K, V ← *ApplyMultiHeadSelfAttention(E)*
    AttentionOutput ← *MultiHeadAttention(Q, K, V)*
    E ← *AddAndNorm(E, AttentionOutput)*
    FFNOutput ← *FeedForwardNetwork(E)*
    E ← *AddAndNorm(E, FFNOutput)*
*End**For***
H ← *E // Final encoded features*
H_bottleneck ← *AggregateFeatures(H)*
End

ALGORITHM 3: DECODER – TRANSFORMER BASED UPSAMPLING

F_up ← *Upsample(H_bottleneck)*
For *each SkipFeature in SkipConnections* ***do***
    F_concat ← *Concatenate(F_up, SkipFeature)*
    Features ← *MultiHeadSelfAtt (F_concat)*
*End**For***
F_refined ← *RefinedFeatures*
S ← *Convolution(F_refined) // Generate segmentation map*
ProbabilityMap ← *Softmax(S) // Apply softmax activation*
For *each Epoch* ***do***
    For *each Batch in TrainingData* ***do***
        Predictions ← *ForwardPass(Batch)*
        Loss ← *CrossEntropy (Predictions, Labels)*
        BackwardPass(Loss)
        UpdateWeights()
    *End**For***
ApplyDropout()
*End**For***
End

Finally, an output is generated through a convolutional layer with a softmax activation that provides detailed probability maps of the tumor classification [8]. The VisionTranscoder

algorithm first normalizes the pixel values of an MRI image. Then, it extracts patches that do not overlap, flattens them, and projects them into high-dimensional embeddings with added positional information [29]. These embeddings are fed through multiple transformer layers using multi-head self-attention and feed-forward networks to compute encoded features. Lastly, these computed features are aggregated to a global representation at the bottleneck step. Upsampling and refining features are done in the decoder through skip connections and additional self-attention. After refining the features, a segmentation map is generated, after which it is converted into a probability map with softmax [31, 32]. Finally, backpropagation is used for computing cross-entropy loss while applying dropout to avoid overfitting. The training process involves multiple epochs with batch updates for optimizing the model's weights.

V. RESULTS

Accuracy measures the overall correctness of a model by calculating the proportion of correct predictions, both positive and negative, out of all predictions made. Precision focuses on the positive predictions and tells you how many of the instances predicted as positive were actually positive. Recall, on the other hand, looks at how well the model identifies all actual positive instances, measuring its9 ability to capture every positive case. The F1 score is a balance between precision and recall, providing a single metric that takes both into account, making it useful when both false positives and false negatives are important to minimize, especially in imbalanced datasets.

The accuracy, precision ,recall and F1 score is calculated using the Eq.(15) [22] to Eq.(18) [22].

$$Accuracy\ (A) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (15)$$

$$Precision\ (P) = \mathrm{TP}\frac{1}{(\mathrm{TP} + \mathrm{FP})} \quad (16)$$

$$Recall\ (R) = TP\frac{1}{(TP + FN)} \quad (17)$$

$$F1 = 2 * \frac{(\mathrm{P} * \mathrm{R})}{(\mathrm{P} + \mathrm{R})} \quad (18)$$

Where TP is true positive , TN is true negative, FP is false positive FN is false negative.

A. PREPROCESSING

FIGURE 2 Image augmentation is a technique used to artificially expand a dataset by applying various transformations to the original images, which helps improve model robustness and generalization. Transformations such as rotation, scaling, flipping, and color adjustment significantly enrich the training data. Rotation alters the image's orientation, allowing the model to recognize objects from different angles, which is useful in cases where objects appear in varying orientations. Scaling adjusts the size of the image, helping the model handle objects of different sizes, making it more adaptable to changes in object proximity or distance. Flipping creates mirror versions of the image, enabling the model to learn from both symmetrical and asymmetrical patterns. Color adjustment, such as changes in brightness, contrast, or saturation, simulates lighting variations, helping the model perform well under different environmental conditions.

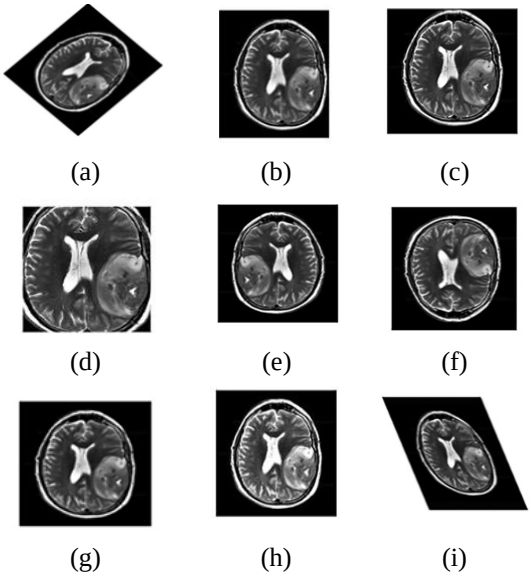

FIGURE 2. (a) to (i) A sample of augmented image from the original image

These transformations increase data diversity, allowing the model to learn more features and improving its performance on unseen data.. FIGURE 3 shows the trend of accuracy of the Vision Transformer (ViT) model during training epochs, illustrating how the model's performance improves over time. Initially, the accuracy is low due to random weight initialization, and the model struggles to make accurate predictions. As the training progresses, the accuracy gradually increases as the model learns to recognize patterns and adjust its parameters. During the middle phase, the accuracy typically experiences steady growth, reflecting the model's ongoing learning process. In later epochs, the accuracy may plateau, indicating that the model has converged to a certain level of performance, and further training may not result in significant improvements. This trend provides valuable insight into the effectiveness of the model's learning and its ability to generalize to new data.. FIGURE 4 complements the accuracy trend shown in FIGURE 3 by depicting the loss of the Vision Transformer (ViT) model over the same epochs. The loss function measures the discrepancy between the model's predictions and the actual ground truth, providing a quantitative measure of how well the model is learning. At the start of training, the loss is typically high, reflecting the large gap between the model's predictions and the true labels. As training progresses, the model adjusts its weights to minimize this discrepancy, causing the loss to decrease over time. A decreasing loss indicates that the model is improving in terms of making more accurate predictions. Similar to the accuracy plot, the loss curve may begin to plateau after several epochs, suggesting that the model has converged and

is no longer making significant improvements. This trend is crucial for evaluating model performance alongside accuracy, as it helps ensure that the model is not just memorizing the training data but also generalizing well to unseen data.

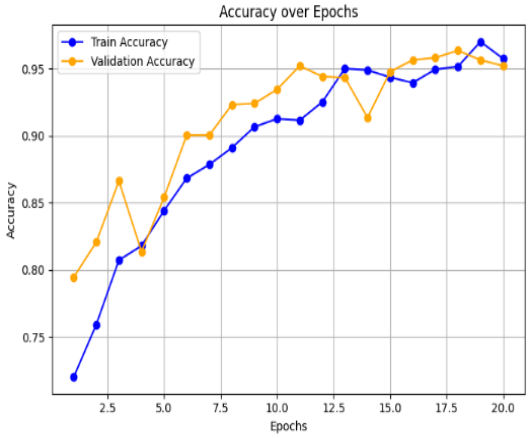

FIGURE 3. An accuracy of vision transformer

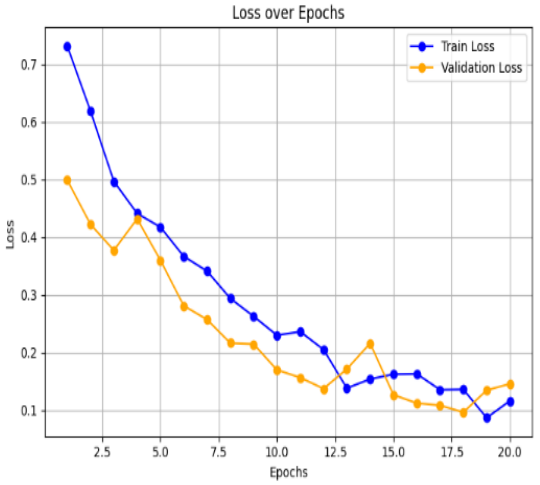

FIGURE 4. A loss of vision transformer

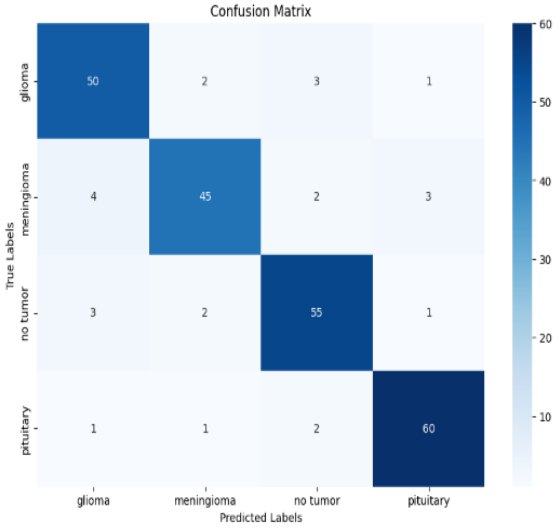

FIGURE 5. Confusion matrix of proposed work

FIGURE 5 shows the confusion matrix of the proposed work, indicating the performance of the model in classifying four classes: glioma, meningioma, no tumor, and pituitary.

TABLE 2
Parameters of the proposed transformer model

S.No	Parameter	Value
1	Batch Size	2, 4, 6, 8, 10, 12, 14
2	Learning Rate	0.001
3	Epochs	20
4	Optimizer	Adam
5	Dropout Rate	0.5
6	Number of Layers	10
7	Hidden Units	128
8	Activation Function	Softmax
9	Weight Initialization	Xavier
10	Loss Function	Cross-Entropy Loss
11	Regularization	L2 Regularization
12	Learning Rate Scheduler	Step Decay
13	Early Stopping	True

The parameters that will be used by the proposed transformer model are shown in TABLE 2. Specifically, the batch sizes that will be used for this model are 2, 4, 6, 8, 10, 12, and 14 so that it works optimally on most of the training tasks. It will make use of the Adam optimization algorithm, which is a pretty efficient optimizer when dealing with sparse gradients. Finally, an initial learning rate of 0.001 will be used for training during 20 epochs. The dropout applied is 0.5 to avoid overfitting, and 10 layers are used with 128 hidden units each with Softmax for multiclass classification. Weight initialization: by Xavier; the loss function is Cross-Entropy Loss for multi-class classification. Regularization through L2 regularization further prevents overfitting. Moreover, it uses the Step Decay scheduler in order to adapt the learning rate, and Early Stopping will stop training once improvements stall, which makes the model more efficient by avoiding extraneous computation.

TABLE 3
A model summary of VisionTranscoder (ViT)

S.No	Layer (Type)	Output Shape	Param #
1	Input Layer	(None, 224, 224, 3)	0
2	Conv2D (Layer 1)	(None, 112, 112, 64)	9,472
3	MaxPooling2D	(None, 56, 56, 64)	0
4	Conv2D (Layer 2)	(None, 28, 28, 128)	73,856
5	MaxPooling2D	(None, 14, 14, 128)	0
6	Flatten	(None, 25088)	0
7	Dense (Encoder - Layer 1)	(None, 512)	12,853,056
8	Dropout	(None, 512)	0
9	Dense (Encoder - Layer 2)	(None, 256)	131,328
10	Encoder Transformer Blocks	(None, 256, 256)	15,000,000

11	Decoder Transformer Blocks	(None, 256, 256)	15,000,000
12	Dense (Decoder - Layer 1)	(None, 512)	131,328
13	Dense (Decoder - Layer 2)	(None, 256)	131,328
14	Output Layer	(None, 4)	1,028

TABLE 3 summarizes the model architecture of VisionTranscoder. It consists of an input layer that takes images of size 224 × 224, and then runs them through two convolutional layers with max pooling, progressively reducing spatial dimensions and increasing feature complexity. Then, there is a Flatten layer that reshapes the feature map into a one-dimensional vector.

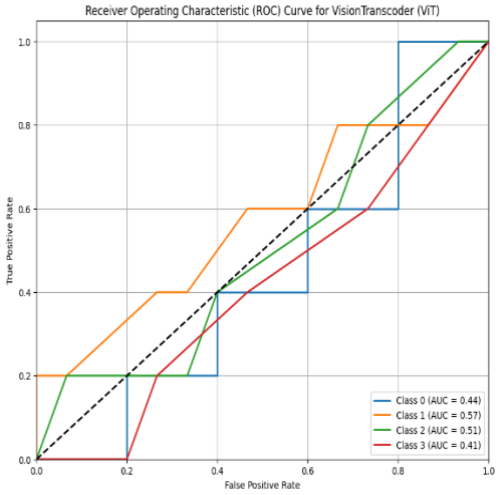

FIGURE 6. RoC curve for VisionTranscoder (ViT)

FIGURE 6 shows RoC curve for the ViT on four classes: glioma, meningioma, no tumor, and pituitary. The plot for this model uses the true positive rate, or sensitivity, versus the false positive rate and specificity at different threshold settings.

TABLE 4
Performance analysis of proposed work with other classifiers

S.No	Model	Accuracy (%)	Loss
1	U-Net	85.5	0.35
2	VGG16	87.2	0.32
3	ResNet	89.0	0.28
4	InceptionV3	88.5	0.30
5	DenseNet	90.3	0.25
6	Data-efficient Image Transformer	91.1	0.22
7	Swin Transformer	89.8	0.27
8	VisionTranscoder (ViT)	98.5	0.05

TABLE 4 demonstrates the performance evaluation of the proposed VisionTranscoder compared with some other classifiers like U-Net, VGG16, ResNet, InceptionV3, DenseNet, Data Efficient Image Transformer, and Swin Transformer. The VisionTranscoder exceeded all models with an accuracy of 98.5% and a loss of 0.05 in both metrics.

The accuracy curve and loss curve for VisionTranscoder are shown in FIGURE 7 and 8, respectively. FIGURE 7 shows that over the epochs, the graph of accuracy is monotonically increasing, thus reflecting superior classification capability. Figure 8 shows a continuous decrease in loss, thus proving that the prediction errors are appropriately minimized and making the training process robust.

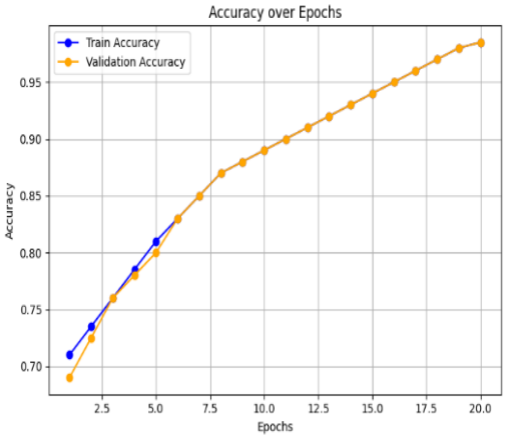

FIGURE 7. An accuracy of proposed VisionTranscoder(ViT)

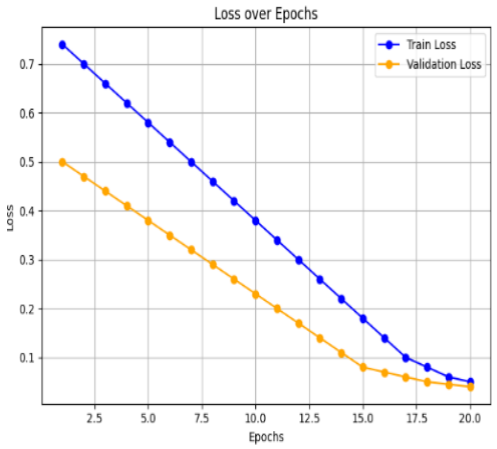

FIGURE 8. A loss of proposed VisionTranscoder(ViT)

TABLE 5
Comparative analysis of models based on metrics (Accuracy, precision, recall, F1 score)

S.No	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	U-Net	85.5	84.0	86.0	85.0
2	VGG16	87.2	85.5	88.0	86.7
3	ResNet	89.0	87.5	89.5	88.5
4	InceptionV3	88.5	86.0	88.2	87.1
5	DenseNet	90.3	88.8	90.5	89.6
6	Data-efficient Image Transformer	91.1	89.5	91.2	90.3
7	Swin Transformer	89.8	87.0	90.0	88.5
8	VisionTranscoder (ViT)	98.5	98.0	98.8	98.4

TABLE 5 compares the different models against some of the key performance metrics such as accuracy, precision, recall,

and the F1 score. In the comparison, VisionTranscoder outperformed all the other models with an accuracy of 98.5%, a precision of 98.0%, a recall of 98.8%, and an F1 score of 98.4%. This clearly depicts its high effectiveness at ensuring highly accurate and reliable classification results.

TABLE 6
Evaluation of Vision Transcoder (ViT) with Different Batch Sizes

S.No	Batch Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	2	97.8	97.5	98.0	97.7
2	4	98.0	97.8	98.2	97.9
3	6	98.1	98.0	98.3	98.1
4	8	98.2	98.1	98.4	98.2
5	10	98.3	98.2	98.5	98.3
6	12	98.4	98.3	98.6	98.4

From TABLE 6, the performance metrics for the ViT model varied from a batch size of 2 to 12. It could be noted that each increase in the batch size showed increased improvements in all the metrics, which included higher rises in accuracy, precision, sensitivity, and F1 score values. In particular, the model performed most optimally at a batch size of 12, showing 98.4% accuracy, 98.3% precision, 98.6% recall, and 98.4% F1 score.

TABLE 7
Response time of various classifiers

S.No	Classifier	Response Time (Minutes)
1	U-Net	25
2	VGG16	30
3	ResNet	28
4	InceptionV3	35
5	DenseNet	32
6	Data-efficient Image Transformer	40
7	Swin Transformer	38
8	VisionTranscoder (ViT)	22

TABLE 7 shows the response time for various classifiers for each model. Among these, VisionTranscoder has a very short response time of only 22 minutes, thus proving that it is quite efficient as compared to the remaining models. This is followed by U-Net, then VGG16 and ResNet with 25, 30, and 28 minutes, respectively. InceptionV3 and DenseNet trail with a response time of 35 and 32 minutes. The response times for Data-efficient Image Transformer and Swin Transformer are the longest at 40 and 38 minutes.

VI. DISCUSSION

From experimental results, the proposed VisionTranscoder model proves to work effectively compared with the traditional models of U-Net, VGG16, ResNet, InceptionV3, DenseNet, and other transformer models (TABLE 8). The maximum recorded classification accuracy shows a value of 98.5% with a loss at 0.05, which depicts the fact that VisionTranscoder substantially minimizes the classification errors while maintaining robustness in the training process.
The proposed model is supported through the application of data augmentation techniques, in addition to raising training data variety and generalization over unseen data. It can be seen from the smooth increase in accuracy and steady reduction in loss in FIGURE 3 and 4 that the model is learning well and is not overfitting. In addition, the confusion matrix in FIGURE 5 shows high performance concerning the correct classification of the four categories that increase the reliability of the model concerning complex medical image segmentation tasks. Compared to other classifier methods, VisionTranscoder outperforms them in aspects not only of accuracy but also of precision, recall, and F1 score. As illustrated in TABLE 5, it achieves a precision of 98.0%, a recall of 98.8%, and an F1 score of 98.4%, significantly outperforming the remaining models compared: U-Net, VGG16, Swin Transformer; it will thus significantly decrease the number of false positives and negatives and much more practical for applications requiring critical medical outputs.

TABLE 8
Comparison of performance among other similar studies

Author / Year	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Chen et al., 2021 [33]	85.24	85.12	83.68	84.39
Cao et al., 2021 [34]	88.76	87.50	88.25	87.87
Xing et al., 2022 [35]	90.15	89.78	89.92	89.85
Zhou et al., 2022 [36]	92.14	91.50	92.00	91.75
Proposed Model	98.5	98.0	98.8	98.4

The other factor by which the efficiency of this model is highlighted is its response time. TABLE 7 and FIGURE 10 show that VisionTranscoder has the shortest response time (22 minutes) in comparison with other models. Therefore, VisionTranscoder is practical to apply in real-time or near-real-time applications in the clinical field. Regarding batch size variations, from TABLE 6, it is clear that for all the increased batch size, it was clear that the model was improving in accuracy, precision, recall, and F1 score in all cases while attaining an optimal batch size at 12. Therefore, this illustrates the fact that the model could be fine-tuned to handle larger batch sizes without reducing its accuracy and this makes it scale up for larger datasets. Therefore, in total, the proposed VisionTranscoder is a very efficient model which provides outstanding performance with respect to correctness, speed, and generalization performance; thus, it can be used for the classification or segmentation of medical images, especially in brain tumor detection. Moreover, the model's ability to treat complex features with transformer blocks and effective optimization techniques, Adam, accompanied by regularization and early stopping, ensures the model's efficiency and reliability.

VII. CONCLUSIONS

Brain tumor detection and classification have limited accuracy with the existing models, while the computational costs usually remain high, offering less than optimum performance on most tumor types. Traditional approaches,

such as U-Net, VGG16, and ResNet, have proved commendable results but have major issues in terms of accuracy and efficiency when handling complex images in medicine. In this regard, the VisionTranscoder model is proposed to handle such problems by considering the state-of-the-art encoder-decoder architectures and Vision Transformer techniques. Experimental results showed that this model performed very well, with an accuracy of 98.5% and a loss of 0.05 without overfitting than other models. Its high accuracy proves that this model works well in recognizing and classifying four types of brain tumors such as glioma, meningioma, no tumor, and pituitary. Though the achieved accuracy of classification for brain tumors in the proposed model VisionTranscoder is up to 98.5%, several drawbacks exist when implementing this model. High computational complexity and considerable training time make it not so applicable to online use or deployment in resource-constrained environments. The model was trained with the application of the dataset specific to a type of certain tumors or medical conditions. It only classifies four kinds of brain tumors. Its scope can be expanded to even greater clinical utility. Future work could include the optimization of the model to make it run faster, an increased scope of the classes to accommodate a greater variety of tumor types or other medical imaging tasks, and improvement in its interpretability with help from methods such as explainability, like Grad-CAM. Apart from that, with the aid of transfer learning for other medical images and multimodal data combination and adapting it to 3D imaging, its wider applicability into medical image analysis is greatly extended.

REFERENCES

[1] S., S., V., S. FACNN: fuzzy-based adaptive convolution neural network for classifying COVID-19 in noisy CXR images. Med BiolEngComput (2024). https://doi.org/10.1007/s11517-024-03107-x

[2] Suganyadevi, S., Pershiya, A.S., Balasamy, K. et al. Deep Learning Based Alzheimer Disease Diagnosis: A Comprehensive Review. SN COMPUT. SCI. 5, 391 (2024). https://doi.org/10.1007/s42979-024-02743-2

[3] Biratu, E. S., Schwenker, F., Ayano, Y. M., &Debelee, T. G. (2021). A survey of brain tumor segmentation and classification algorithms. Journal of Imaging, 7(9), 179.

[4] Rao, C. S., &Karunakara, K. (2021). A comprehensive review on brain tumor segmentation and classification of MRI images. Multimedia Tools and Applications, 80(12), 17611-17643.

[5] Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., & Liu, Y. (2023). Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Information Fusion, 91, 376-387.

[6] Shamia, D., Balasamy, K., Suganyadevi, S.: A secure framework for medical image by integrating watermarking and encryption through fuzzy based ROI selection. J. Intell. Fuzzy Syst. 44(5), 7449–7457 (2023)

[7] Naser, M. A., &Deen, M. J. (2020). Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Computers in biology and medicine, 121, 103758.

[8] Gómez-Guzmán, M. A., Jiménez-Beristaín, L., García-Guerrero, E. E., López-Bonilla, O. R., Tamayo-Perez, U. J., Esqueda-Elizondo, J. J., ... &Inzunza-González, E. (2023). Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks. Electronics, 12(4), 955.

[9] Bal, A., Banerjee, M., Chakrabarti, A., & Sharma, P. (2022). MRI brain tumor segmentation and analysis using rough-fuzzy c-means and shape based properties. Journal of King Saud University-Computer and Information Sciences, 34(2), 115-133.

[10] S. Suganyadevi, V. Seethalakshmi, K. Balasamy and N. Vidhya, "Deep learning in Covid-19 detection and diagnosis using CXR images: challenges and perspectives", Digital Twin Technologies for Healthcare, vol. 4, no. 046, pp. 163, 2023.

[11] Kumar, K. A., Prasad, A. Y., &Metan, J. (2022). A hybrid deep CNN-Cov-19-Res-Net Transfer learning architype for an enhanced Brain tumor Detection and Classification scheme in medical image processing. Biomedical Signal Processing and Control, 76, 103631.

[12] Krishnasamy, B., Balakrishnan, M., Christopher, A. (2021). A GeneticAlgorithm Based Medical Image Watermarking for ImprovingRobustness and Fidelity in Wavelet Domain. In: Satapathy, S., Zhang,YD., Bhateja, V., Majhi, R. (eds) Intelligent Data Engineering andAnalytics. Advances in Intelligent Systems and Computing, vol 1177.Springer, Singapore. https://doi.org/10.1007/978-981-15-5679-1_27.

[13] Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., ... & Ye, X. (2018). Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels. Computer methods and programs in biomedicine, 157, 69-84.

[14] Dataset collection:kagge repository-https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[15] Selvapandian, A., &Manivannan, K. (2018). Fusion based glioma brain tumor detection and segmentation using ANFIS classification. Computer methods and programs in biomedicine, 166, 33-38.

[16] Suganyadevi Sellappan, A Anand Shiny Pershiy, Finney Daniel Shadrach, Krishnasamy. Balasamy, Karra. Renu and UmaamaheshvariAnnamalai, "A survey of Alzheimer's disease diagnosis using deep learning approaches", Journal of Autonomous Intelligence, vol. 7, no. 3, 2024.

[17] Abdel-Maksoud, E., Elmogy, M., & Al-Awadi, R. (2015). Brain tumor segmentation based on a hybrid clustering technique. Egyptian Informatics Journal, 16(1), 71-81.

[18] Rehman, Z. U., Naqvi, S. S., Khan, T. M., Khan, M. A., & Bashir, T. (2019). Fully automated multi-parametric brain tumour segmentation using superpixel based classification. Expert systems with applications, 118, 598-613.

[19] Ranjbarzadeh, R., BagherianKasgari, A., JafarzadehGhoushchi, S., Anari, S., Naseri, M., &Bendechache, M. (2021). Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Scientific Reports, 11(1), 1-17.

[20] Sharif, M. I., Li, J. P., Khan, M. A., &Saleem, M. A. (2020). Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images. Pattern Recognition Letters, 129, 181-189.

[21] Allah, A. M. G., Sarhan, A. M., &Elshennawy, N. M. (2023). Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information. Expert Systems with Applications, 213, 118833.

[22] Suganyadevi, S., Pershiya, A.S., Balasamy, K. et al. Deep Learning Based Alzheimer Disease Diagnosis: A Comprehensive Review. SN COMPUT. SCI. 5, 391 (2024). https://doi.org/10.1007/s42979-024-02743-2

[23] Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., &Jhanjhi, N. Z. (2022). A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. Irbm, 43(4), 290-299.

[24] Hussain, S., Anwar, S. M., & Majid, M. (2018). Segmentation of glioma tumors in brain using deep convolutional neural network. Neurocomputing, 282, 248-261.

[25] Chen, S., Ding, C., & Liu, M. (2019). Dual-force convolutional neural networks for accurate brain tumor segmentation. Pattern Recognition, 88, 90-100.

[26] Balasamy, K., Suganyadevi, S. Multi-dimensional fuzzy based diabetic retinopathy detection in retinal images through deep CNN method. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19798-1

[27] Ghassemi, N., Shoeibi, A., & Rouhani, M. (2020). Deep neural network with generative adversarial networks pre-training for brain tumor

classification based on MR images. Biomedical Signal Processing and Control, 57, 101678.
[28] Öksüz, C., Urhan, O., &Güllü, M. K. (2022). Brain tumor classification using the fused features extracted from expanded tumor region. Biomedical Signal Processing and Control, 72, 103356.
[29] Balasamy, K., Seethalakshmi, V. &Suganyadevi, S. Medical Image Analysis Through Deep Learning Techniques: A Comprehensive Survey. Wireless PersCommun 137, 1685–1714 (2024). https://doi.org/10.1007/s11277-024-11428-1
[30] Sompong, C., &Wongthanavasu, S. (2017). An efficient brain tumor segmentation based on cellular automata and improved tumor-cut algorithm. Expert Systems with Applications, 72, 231-244.
[31] S., S., V., S. FACNN: fuzzy-based adaptive convolution neural network for classifying COVID-19 in noisy CXR images. Med BiolEngComput 62, 2893–2909 (2024). https://doi.org/10.1007/s11517-024-03107-x.
[32] Suganyadevi, S., Seethalakshmi, V. Deep recurrent learning based qualified sequence segment analytical model (QS2AM) for infectious disease detection using CT images. Evolving Systems 15, 505–521 (2024). https://doi.org/10.1007/s12530-023-09554-5.
[33] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., & Zhou, S. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. ArXiv preprint arXiv:2102.04306. Available at: https://arxiv.org/abs/2102.04306
[34] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Swin Transformer for Medical Image Segmentation. ArXiv preprint arXiv:2105.05537. Available at: https://arxiv.org/abs/2105.05537
[35] Xing, W., Wang, F., Liu, X., & Li, Z. (2022). CS-Unet: A Compact Skip-Connected UNet for Medical Image Segmentation. ArXiv preprint arXiv:2210.08066. Available at: https://arxiv.org/abs/2210.08066
[36] Zhou, Y., Li, J., Wang, X., Feng, Y., & Zhang, Y. (2022). MedFormer: A Data-scalable Transformer for Medical Image Segmentation. ArXiv preprint arXiv:2203.00131. Available at: https://arxiv.org/abs/2203.00131

AUTHORS BIOGRAPHY

Saroj Bala is working as an Associate Professor in the Department of Master of Computer Applications, Ajay Kumar Garg Engineering College, Ghaziabad, India. She has completed her Ph.D. from Shobhit University, Meerut, MCA from Punjabi University, Patiala, and B.Sc.(CS) from Kurukshetra University, Kurukshetra. She has 25 years of teaching experience. Her research interests include swarm intelligence, machine learning, deep learning, data science, cyber security, and image processing. She has attended several seminars, workshops, and conferences. She has published many research papers in national and international journals.Email: saroj.chhokar@gmail.com

Kumud Arora is working as a Professor with CSE- Artificial Intelligence & Machine Learning at Inderprastha Engineering College, Ghaziabad, India. She has completed her Ph.D. from Banasthali Vidyapith, Banasthali. She has 23 years of teaching experience. Her research interests include machine learning, deep learning applications to agriculture, cyber forensics. She has attended several seminars, workshops, and conferences. She has published many research papers in national and international journals. She is also the recipient of National Programme on Technology Enhanced Learning (NPTEL) Discipline Star for 2023. Email: Kumud.kundu@ipec.org.in

Jeevitha R is an Assistant Professor in the Department of CSE at KPR Institute of Engineering and Technology, Coimbatore. She earned her B. E degree in Computer Science and Engineering from KPR Institute of Engineering and Technology in 2016, an M. E in Computer Science and Engineering from KPR Institute of Engineering and Technology in 2018.She has published approximately 10 papers in international and National and conferences. Her areas of interest include Machine Learning, Image Processing and Blockchain Technologies. Email: jeedhar95@gmail.com

Rini Chowdhury is Sub Divisional Engineer presently working in the department of IT Project Circle in Bharat Sanchar Nigam Limited (BSNL), Kolkata, West Bengal working as a key member of reports team on Oracle based DB and application platform. She received her B.E degree in Electronics and Telecommunication stream in the year of 2009 from Jadavpur University, Kolkata, India. She has done her B.E. research work at the IC Design and Fabrication Centre, Department of Electronics & Telecommunication Engineering, Jadavpur University. Her undergraduate research interest included medical diagnosis using soft computing techniques. She has published 3 IEEE international conference papers in this domain and now extending her interest in image processing domain and has published 1 IEEE international conference paper in it.Email: chowdhury.rini@gmail.com

Prashant Kumar is Junior Engineer presently working in the department of IT Project Circle in Bharat Sanchar Nigam Limited (BSNL), Kolkata, West Bengal working in the capacity of Oracle based application developer on PL/SQL platform He received his B. Tech. degree in Electronics and Telecommunication stream in the year of 2013 from KIIT University, Bhubaneswar, India. He has also worked as System Engineer in Tata Consultancy Services Limited (TCS), Lucknow, Uttar Pradesh. He has published 2 IEEE international conference papers. His research interest includes image processing using soft computing techniques.Email: k21prashant@gmail.com

Dr.C.Shobana Nageswari, is Associate Professor in the Department of Electronics and Communication Engineering at R.M.D Engineering College where she has been a faculty member since June 2007. She obtained B.E in Instrumentation and Control Engineering in the year 2000 from Adhiyamaan College of Engineering, and M.E. in Applied Electronics Degree in the year 2007 at College of Engineering, Anna University. She has completed her Ph.D under Anna University in the area of Medical Image Processing in the year 2019.She has 20 years of teaching experience and has guided many B.E. projects. Her areas of interest include Soft computing,Bio Medical image processing and Healthcare, She presented and published 20 papers in the international journals and conferences. She published two patents.She is a life time member of IETE and ISTE.Email:shobananageswari79@gmail.com