

RESEARCH ARTICLE

LCDEiT: A Linear Complexity Data-Efficient Image Transformer for MRI Brain Tumor Classification

GAZI JANNATUL FERDOUS¹, KHALEDA AKHTER SATHI¹,MD. AZAD HOSSAIN¹, (Member, IEEE),MOHAMMED MOSHIUL HOQUE², (Senior Member, IEEE),AND M. ALI AKBER DEWAN³, (Member, IEEE)¹Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh²Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh³School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Athabasca, AB T9S 3A3, Canada

Corresponding authors: Khaleda Akhter Sathi (sathi.ete@cuet.ac.bd), Md. Azad Hossain (azad@cuet.ac.bd), and Mohammed Moshuiul Hoque (moshiul_240@cuet.ac.bd)

ABSTRACT Current deep learning-assisted brain tumor classification models sustain inductive bias and parameter dependency problems for extracting texture-based image information. Thereby concerning these problems, the recent development of the vision transformer model has substituted the DL model for classification tasks. However, the high performance of the vision transformer model depends on a large-scale dataset as well as self-attention calculations between the number of image patches which result in a quadratic computational complexity. To address these problems, the vision transformer must be data-efficient to be well-trained with a limited amount of data, and the computational complexity must be linear with the number of image patches. Consequently, this paper presents a novel linear-complexity data-efficient image transformer called LCDEiT for training with small-size datasets by using a teacher-student strategy and linear computational complexity concerning the number of patches using an external attention mechanism. The teacher model comprised a custom gated-pooled convolutional neural network to provide knowledge to the transformer-based student model for the classification of MRI brain tumors. The average classification accuracy and F1-score for two benchmark datasets including Figshare and BraTS-21 are found 98.11% and 97.86% and 93.69% and 93.68% respectively. The results indicate that the proposed model could have a great impact on medical imaging-based diagnosis where data availability and faster computations are the main concern.

INDEX TERMS Brain tumor, classification, external attention, MRI, transformer.

I. INTRODUCTION

The mortality rate due to brain cancer can be minimized by detecting brain tumors of the specific class in the earlier stage. Several imaging techniques such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) have an impact on the earlier detection of brain cancer. Among these, MRI has mostly used imaging techniques in the medical field [1], [2]. The

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan¹.

identification of brain tumors from these imaging techniques can lead to a false diagnosis which may cause a threat to life [3]. Therefore, the utilization of an automated system for quickly identifying the correct brain tumor class can aid a significant role in traditional imaging techniques [4]. Currently, automated systems are developed based on advanced technology such as machine learning (ML) [5], [6], [7], [8], [9] and deep learning (DL) [10], [11], [12], [13], [14] algorithms to identify the brain tumor classes precisely. However, the ML techniques have limitations to process image-type datasets as handcrafted feature extraction is needed before data

processing [15], [16]. On the other side, DL techniques have the advantages to extract meaningful features automatically before classification [17].

In earlier studies, convolutional architecture dominates the DL field for computer vision tasks such as classification, segmentation, object detection, and recognition. The Pre-trained convolutional neural network (CNN) i.e., residual neural network (ResNet) has outperformed the other convolutional network architecture [18] for the classification task. The CNN models are most case architecture-specific due to depending on the parameter and training procedures. Moreover, it focuses on texture information based on the assumption to generate output using locality and weight-sharing features which leads to inductive bias. Therefore, the vision transformer (ViT) model is developed as a replacement for the CNN model due to having better computational efficiency and scalability. The ViT model incorporates a self-attention-based core transformer model for finding the relations between non-overlapping patches of the image. Then, the parallelly executed multiple self-attention task called multi-head self-attention pays attention to a particular image feature for classifying the images into their actual class using a fully connected dense layer [19]. However, the performance of ViT model is limited to two concerns. One is a large-sized dataset requirement for optimal model accuracy. The other one is quadratic computational complexity w.r.t image size due to the employment of a self-attention mechanism. From these two concerns, the advanced ViT model called the Swin transformer focus on complexity concern which is linear to the image size. The computation is performed by calculating non-overlapping window-based local self-attention. Where the complexity is linear by computing self-attention through a shifted window between consecutive layers [20]. However, the requirement of sufficient data for the high performance of the Swin transformer is still a problem to be looked up. On the contrary, the Data-efficient image transformer (DEiT) is only capable of handling small-sized data with the help of a teacher model. A distillation token helps the student model to adapt the knowledge of the teacher model through attention [21]. Where the multi-head self-attention computes the relation between the patches which leads to quadratic complexity concerning the number of patches. For this reason, the self-attention technique can be replaced by an external attention mechanism that is based on a learnable memory unit to reduce the quadratic computational complexity of classification tasks [22]. Therefore, to overcome the two limitations presented in the ViT model such as enormous dataset requirements and quadratic computation, a model needs to be data efficient and computation needs to be linear to image size without compromising the model accuracy. The primary contributions of this work are outlined as follows:

- A linear complexity data-efficient image transformer (LCDEiT) is developed to classify brain tumors that can provide a great impact on future medical imaging fields.

- A custom-gated pooled CNN network is employed as a teacher model to distill knowledge to a transformer-based student model for providing data efficiency by reducing the requirement of a large dataset and contributing to calculating the cross-entropy loss.
- A multi-head external attention mechanism is introduced to provide a linear computation w.r.t number of patches which ultimately reduces model training parameters and time without compromising the classification accuracy.

The residual part of the paper is organized as follows: Section II presents an overview of the related research on the classification task. A detailed explanation of the proposed methodology is presented in section III. The description of the dataset and evaluation matrices are outlined in Section IV. The result is described in detail in Section V with some performance measurements. A comparative analysis with the state-of-the-art models is presented in discussion Section VI. Finally, the conclusion and future directions of the paper are drawn in Section VII.

II. RELATED WORK

Earlier studies commenced with the ML algorithm as a base model for computer vision tasks. For instance, M. A. et al. [8] performed a gray-level co-occurrence matrix (GLCM) for statistical feature extraction and discrete wavelet transform (DWT) for brain tumor segmentation which augment the performance and shrivel the complexity. The noise emerged due to segmentation is eradicated by morphological operation and then classification is performed by a support vector machine (SVM) classifier. Moreover, Prabhpreet et al. [9] proposed an MRI brain tumor detection technique including several stages such as tumor segmentation, and statistical feature extraction followed by binary classification into benign and malignant. A modified medial filtering and multi-vector segmentation method support the SVM classifier for tumor classification. For the ML-based classifier, the generation of statistical features from raw images is handcrafted and user-specific which results in degrading model performance. Thereby, the later studies treated CNN as the standard framework for computer vision tasks due to the ability to extract important features automatically from the raw images. Ghosal et al. [18] employed a squeeze and excitation ResNet model based on CNN for the image classification task. In addition, the utilization of zero centering and intensity normalization provided smooth variation in the intensity which increases the effectiveness of the classification task. However, the CNN-based architecture is restricted to local features and the model performance is affected by the inductive biasing problem. Therefore, transformer-based architectures are developed to extract global information from the input images. In one study, Dosovitskiy et al. [19] proposed a ViT model to act on several computer vision tasks including classification, segmentation, detection, and recognition. The individual tasks initiated with making patches from images and feeding the projection of images into the transformer encoder. Then, the output of

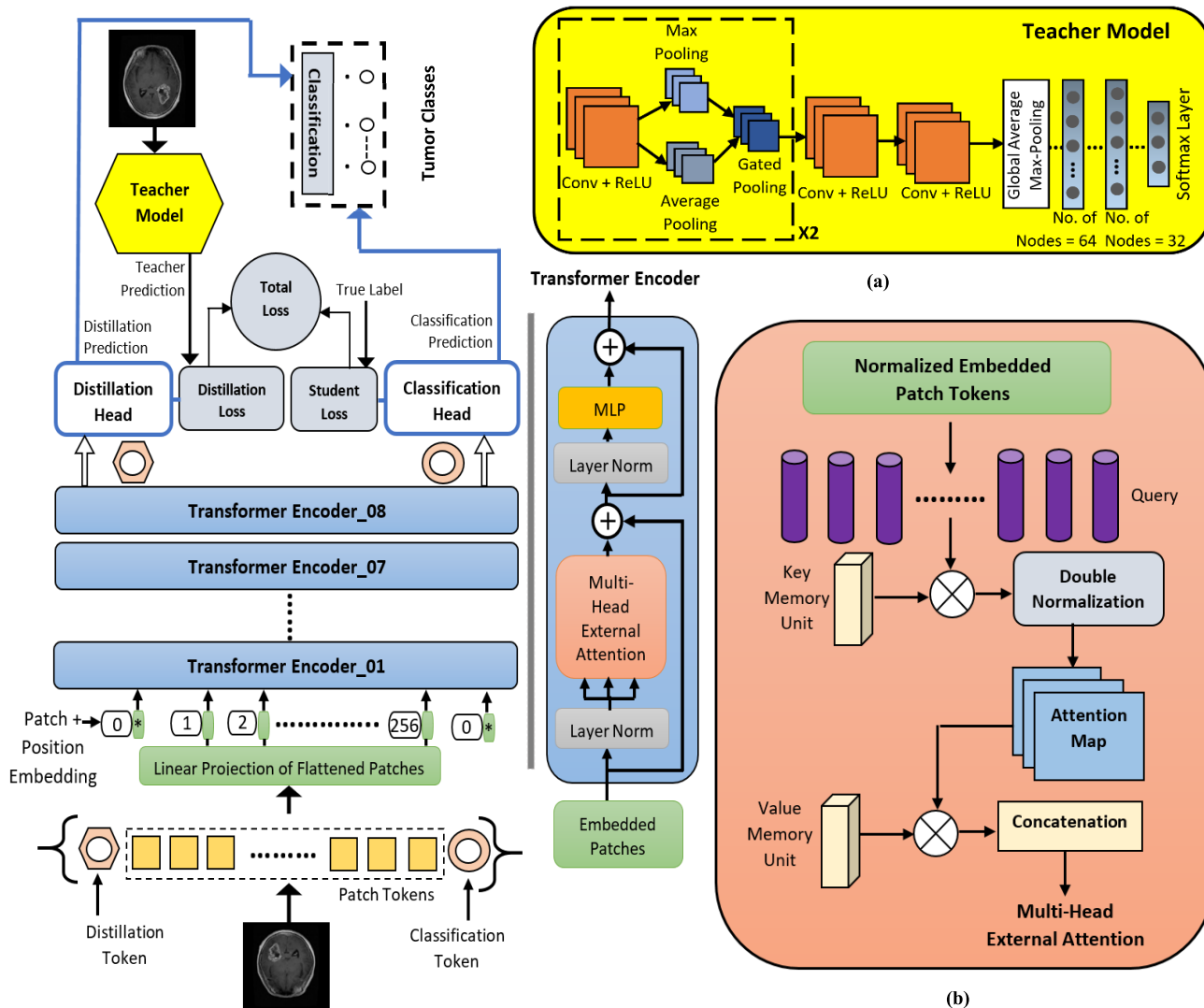


FIGURE 1. Illustration of proposed linear complexity data efficient image transformer (LCDEiT). (a) Gated-pooled based customized CNN teacher model, (b) Multi-head external-attention mechanism for transformer based student model.

the encoder was applied for specific task prediction. The requirement of a vast dataset for training is one of the main bottlenecks of this work. Another limitation such as quadratic complexity w.r.t image size is present in the work due to the multi-head self-attention mechanism. To pay attention only to the complexity concern, Liu et al. [20] introduced a Swin transformer where self-attention was computed between non-overlapping windows, which results in linear complexity to image size. However, the requirement of an enormous dataset is still a problem for the superior performance of the Swin transformer. Moreover, to focus only on the large dataset problem, Touvron et al. [21] proposed a model called DEiT that relied on a distillation token to make a model well-trained with insufficient data. The pre-trained RegNetY-16GF-based teacher model distilled knowledge to the transformer-based student model through a distillation token. Where the student

model employed a multi-head self-attention mechanism for final classification that leads to computational complexity quadratic in nature. In another study, Tolstikhin et al. [23] developed MLP-Mixer architecture for computer vision tasks that eliminated convolution as well as the self-attention mechanism. The work was mainly based on multi-layer perception for token mixing and channel mixing separately. Then linear layer was used for the final classification. But the degradation of accuracy compared to ViT and the requirement of large data is still an issue for this work. Moreover, Wang et al. [24] proposed a pyramid ViT that employed a linear-complexity attention layer by spatial reduction attention (SRA) and tokenized images with overlapping patch embedding to extract local continuity of information. The utilization of SRA makes the computational complexity linear, however, the vast amount of data is still necessary to

get optimum results. For this reason, Lee et al. [25] utilized shifted patch tokenization (SPT) model to embed more spatial information in the visual token where the spatially shifted images were concatenated with the input image. The Local self-attention (LSA) mechanism sharpens the distribution of attention scores to reduce smoothing problems but the complexity is still quadratic with the image size. Moreover, Trockman et al. [26] designed a conv-mixer model that used convolution for mixing spatial and channel dimensions. Where depth-wise convolution mixed spatial location and after that pointwise convolution mixed channel location to increase data efficiency. This model is designed with compromising the accuracy performance of small-sized datasets. Similarly, the Shift-ViT model introduced by Wang et al. [27] replaced attention with zero parameter shift operation. The model classification was performed by linear layer. The elimination of attention operation in the model results in no complexity concerns but the vast amount of data is still a problem for superior model performance. On the other hand, Zhang et al. [28] used a transformer for covid-19 diagnosis from the chest CT images. After the segmentation of lung images with UNet, Swin transformer is used for feature extraction. However, this model also suffers enormous data requirement problems related to the Swin transformer. To concern this limitation, Zhiqin et al. [29] employed shifted patch tokenization on swin transformer for a specific task of brain tumor segmentation by fusing deep semantics and edge information of multimodal MRI. Despite performing linearly complex able feature extraction using swin transformer, the edge feature extraction is CNN sensitive which may lead to an inductive biasing problem. In another work, a spatial-channel feature preserving vision transformer (SCViT) proposed by Pengyuan et al. [30] extracted long-range dependencies between features and considered the contribution of the different channels in the classification by computing lightweight channel attention. This version of ViT suffers from both limitations such as quadratic complexity and vast data requirement. Similarly, Bazi et al. [31] used a ViT for remote sensing image classification with several data augmentation techniques such as cutmix, cutout, and mixup to get sufficient data to train. Without compromising the accuracy, half of the layers from the model are pruned to reduce parameters and complexity. Moreover, Wang et al. [32] proposed vision transformer-plus (ViT-P) architecture which made a balance between category imbalances by applying deep convolutional generative adversarial networks (DCGAN). Then, channel attention correlated with different channels and obtains important features of each channel for the classification task. The performance of the architectures used in works [31] and [32] is limited by the core two limitations of the ViT model.

In summary, the existing transformer-based classification model suffers from the calculation of self-attention leads to computational complexity quadratic to the number of pixels and the requirement of an enormous dataset for superior classification results. Therefore, the utilization of an external

attention-based transformer model as a student model and a customized gated-pooled-based CNN model as a teacher model can overcome the deficiency of the state-of-the-art classification models.

III. LCDEiT FRAMEWORK

Figure 1 illustrates the proposed LCDEiT framework for MRI brain tumor classification. Where a teacher-student strategy allows the student model to learn through external attention and distill knowledge from the teacher model. A gated-pooled-based customized CNN model is utilized as a teacher model that provides data efficiency flexibility to the student model for classification on small-sized datasets. The customized gated pooled CNN is designed to generalize the model based on the data fed into it. The teacher model contributes to calculating and minimizing the total cross-entropy loss in the overall LCDEiT architecture. Moreover, an external attention-based transformer model is employed as a student model which calculates the attention between patches linearly for final classification. Additionally, a descriptive explanation of the overall classification procedure is given in the subsequent subsections.

A. IMAGE PROCESSING

Initially, the raw images of size (512×512) are resized into (32×32) before patch patching. As the transformer process sequence of image patch tokens, a fixed-size input image is initially converted to non-overlapping patches of fixed size. The raw image, I with dimension $(H \times W) \in \mathbb{R}^{32}$ and the patch with resolution $(P \times P) \in \mathbb{R}^2$ generates a total number of $N \in (H \times W)/P^2 \in \mathbb{R}^{256}$ patches. Equation (1) presents the formulation of patch matrix, $I_{Patching}$ from the raw image. Then, the $I_{Patching}$ are projected to a feature vector using a linear layer that conserves a fixed dimension, $D \in \mathbb{R}^{64}$ which results in a patch token, I_p . After that, position embedding, E_{pos} is added to each patch token, I_p to retain position information that formulated projected output, $I_{Projection}$.

$$I_{Patching} = \begin{bmatrix} [P_1] & \cdots & [P_{16}] \\ \vdots & \ddots & \vdots \\ [P_{241}] & \cdots & [P_{256}] \end{bmatrix} \quad (1)$$

$$I_{Projection} = \begin{bmatrix} [I_{p_1}^1 \dots I_{p_1}^{64}] + E_{pos1} \\ \vdots \\ [I_{p_{16}}^1 \dots I_{p_{16}}^{64}] + E_{pos16} \\ \vdots \\ [I_{p_{241}}^1 \dots I_{p_{241}}^{64}] + E_{pos241} \\ \vdots \\ [I_{p_{256}}^1 \dots I_{p_{256}}^{64}] + E_{pos256} \end{bmatrix} \quad (2)$$

B. CLASS TOKEN

A trainable classification token, T_c is prepended to the generated $I_{Projection}$ to feed into the transformer encoder. Where the token is applied to the classification head for tumor class

prediction. The transformer encoder makes an interrelation between patch tokens, I_p and classification tokens, T_c through external attention with a dimension of $D \in \mathbb{R}^{64}$, but the classification token is only responsible for predicting the final output. Moreover, the classification token calculates the student loss in the training stage and the final class in the testing stage.

C. DISTILLATION TOKEN

Another token called the distillation token, T_d is added to the $I_{Projection}$ and T_c to establish a relationship through external attention in the transformer encoder. The distillation tokens prediction is contributed to calculating distillation loss in the training stage and the average of prediction from the classification and distillation token is used at the testing stage for final classification.

$$T_{in} = [T_c; I_{Projection}; T_d] \quad (3)$$

The input of the transformer encoder, T_{in} are formulated with a dimension of $\{N+2 \text{ (two tokens)} \times D\} \in \mathbb{R}^{258 \times 64}$ by adding two class tokens such as T_d and T_c with $I_{Projection}$ simultaneously. The classification and distillation tokens are initialized by zero having dimension, $D \in \mathbb{R}^{64}$ and updated during training.

D. TRANSFORMER ENCODER

The transformer encoder, T_E consists of multi-head external attention (EA) followed by multi-layer perception (MLP). The details of the EA are described later in the later subsection. The MLP block consists of two fully connected linear layers with an activation function of the gaussian error linear unit (GELU). The number of nodes in two fully connected layers is equal to the projected feature dimension, $D \in \mathbb{R}^{64}$. Moreover, a skip connection is maintained on both EA and MLP to ensure feature reusability and solve the degradation problem. The normalization layer in both EA and MLP blocks normalizes the summed input to reduce dependencies between instances. In this work, a stack of eight identical transformer encoders is used having 4-head EA and an MLP block of [64, 64] units. Moreover, TABLE 1 specified all the required parameters of the transformer-based student model in the proposed LCDEiT architecture. Additionally, TABLE 2 depicts the shape and number of parameters of several blocks presented in the student model. The image is patching with dimension (2×2) , which results in $N \in \mathbb{R}^{256}$ patches per image. Then, the patches are projected to a fixed feature dimension, $D \in \mathbb{R}^{64}$. Therefore, the shape of the projection block is $(256, 64)$. Then, the T_E allows patch token with two extra tokens such as classification and distillation tokens which cause the shape of $(258, 64)$. Both the shape and parameters of T_E is retained the same throughout the eight stacked transformer encoders that allow controlling the parameter count of the overall transformer-based student model.

E. EXTERNAL ATTENTION

The core function of the transformer encoder, T_E is based on the EA mechanism that provides linear complexity to the proposed method. Figure 2 illustrated a visualization of the complexity assessment of self-attention (SA) and external attention (EA). In the traditional SA mechanism, query (Q), key (K), and value (V) vector is generated from each patch. The matrix multiplication of Q and K results QK^T . Then, normalization of this output is again matrix multiplied with V. Calculating attention in this way leads to the requirement of N operation for a single patch where N is the total number of patches in an image. Therefore, the completion of attention calculation for the whole image requires N^2 operation. This functionality is depicted in FIGURE 2(a) where an image consists of $N \in \mathbb{R}^{256}$ patches and calculation of attention for each patch, SA_{Patch_1} need $N \in \mathbb{R}^{256}$ operations that lead to the quadratic computational complexity $O(N^2)$ to calculate whole image attention.

$$SA_{Patch_1} = \sum_{N=1}^{256} Norm(Q_1 K_N^T) V_N \quad (4)$$

On the contrary, the mechanism of EA computes the pixel-wise relation between patches of images and memory units. Two learnable parameters that are independent of input features are introduced externally as a key memory unit, M_k and value memory unit, M_v . Only the query vectors generated from normalized patch tokens, $T_{in-norm}$ leads to a reduction of the input-dependent variable. It allows an increment of the robustness of the

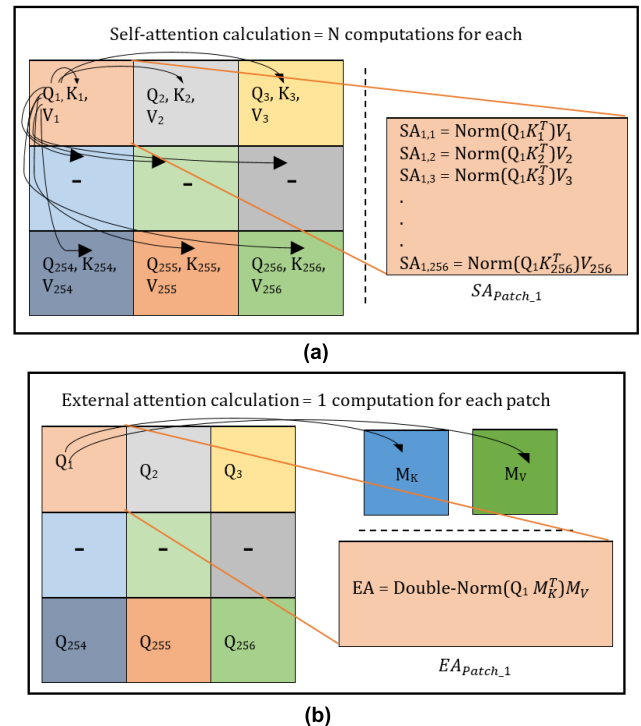


FIGURE 2. Visualization of complexity assessment of self-attention (SA) and external attention (EA) mechanism.

model as compared to self-attention where all three vectors (Q, K, V) are generated from the patch. An illustration of EA mechanism is presented in FIGURE 2(b) where an image having $N \in \mathbb{R}^{256}$ patches and attention calculation for a single patch, EA_{Patch_1} need only one operation which helps to achieve linear computational complexity $O(N)$ for whole image attention over SA. A generalized version of the whole EA process is depicted in (5).

$$EA_{Patch_1} = Double - Norm \left(Q_1 M_k^T \right) M_v \quad (5)$$

The computation of correlation between input patches and shared memory is utilized by employing only two linear layers and two normalization layers. The matrix multiplication of the self-query vector and learnable key memory unit, M_k is double normalized. Then this attention map, A_m is again matrix multiplied with the learnable value memory unit, M_v which generates external attention output, T_{out} (see FIGURE 1(b)). In the transformer encoder, the normalization of T_{in} is performed and the produced normalized patch tokens, $T_{in-norm}$ is fed into a multi-head external attention block. For both key and value memory units, the dimension is chosen (M_k, M_v) $\in \mathbb{R}^{16}$.

$$A_m = Double - Norm \left(T_{in-norm} M_k^T \right) \quad (6)$$

$$T_{out} = A_m M_v \quad (7)$$

TABLE 1. Parameters in the student model of proposed LCDEiT architecture.

Parameter	Value
Image size, (H, W)	32, 32
Patch size, P	2
Feature dimension, D	64
Number of head, H	4
Number of transformer encoders	8
MLP units	[64, 64]

TABLE 2. Properties of blocks in the student model of LCDEiT architecture. (D1 and D2 indicate required parameters for Figshare and BraTS-21 datasets respectively).

Block	Shape	Parameter	
Projection	(256,64)	320	
Transformer Encoder_01	(258,64)	42208	
Transformer Encoder_02	(258,64)	42208	
Transformer Encoder_03	(258,64)	42208	
Transformer Encoder_04	(258,64)	42208	
Transformer Encoder_05	(258,64)	42208	
Transformer Encoder_06	(258,64)	42208	
Transformer Encoder_07	(258,64)	42208	
Transformer Encoder_08	(258,64)	42208	
Layer Normalization	--	128	
Classification Head	--	195 (D1)	260 (D2)
Distillation Head	--	195 (D1)	260 (D2)
		Total =	Total =
		338502	338632
		(D1)	(D2)

Double normalization technique utilizes normalization technique twice by separately normalizing both columns and

rows to handle the sensitivity of the input features scale. At first, single normalization is applied to the matrix multiplied output of query from patches and key memory unit. The exponent of each element is divided by a row-wise summation of the exponential of each element where $R \in \mathbb{R}^{258}$ as patch tokens with additional two tokens make a total of 258 rows. Then the output of single normalization, S_{norm} is again normalized by dividing with the column-wise summation of all single normalized vectors which is referred to as a double normalized vector, D_{norm} where $C \in \mathbb{R}^{64}$ as the feature dimension is 64. The equation of double normalization is depicted as follows:

$$S_{norm} = \exp \left(T_{in-norm} M_k^T \right) / \sum_{R=258} \{ \exp \left(T_{in-norm} M_k^T \right) \}_R \quad (8)$$

$$D_{norm} = S_{norm} / \sum_{C=64} \{ S_{norm} \}_C \quad (9)$$

The multi-head EA is performed by repeating the EA computation multiple times in parallel, each of these is referred to as attention head, h. This process leads to a boost in performance by extending the learning capability of the model to capture different aspects of the relation between patches. Query vector from each patch is transformed independently into H linearly projected query vectors using dense layers where H refers to the total number of heads. These projected query vectors, external key, and value memory units are embedded to calculate the attention score H times in parallel. Then concatenation of the H attention score for each head, (h_1, \dots, h_H) is transformed with another linear projection matrix, W_o which refers to the multi-head external attention output, T_{multi_head} . This transformation matrix helps to make the dimension of input and output consistent.

$$T_{multi_head} = Concatenation (h_1, h_2, \dots, h_H) W_o \quad (10)$$

In this work, 4-head external attention is computed to extract the relation between patches. Four query vectors are extracted from a normalized patch token and after double normalization, four attention maps are found.

F. TEACHER MODEL

The teacher-student strategy works based on a knowledge distillation framework. Knowledge distillation is a model compression technique where a heavy-weight complex model transfers knowledge to a lightweight student model. A strong image classifier such as a convolutional neural network or transformer can be utilized as a teacher model. Earlier studies imply that the ConvNet teacher model performs better than the transformer-based teacher model. In this work, the core idea of knowledge distillation is utilized in a slightly different way as one of our concerns is to reduce complexity. Instead of taking a heavy-weight model to distill knowledge into a student model, a customized lightweight gated pooled CNN is utilized which can learn complex patterns [40] from the data fed into it in replacement of RegNetY-16GF as used in the traditional teacher model. The less complexity property of the teacher model provides fewer parametric quantity that leads

Algorithm 1 External Attention Based MRI Brain Tumor Multi-Classification**Input:** Input Image I ,True Label of Image y .**Output:** Predicted Class of Image P_F .**#Student Model**

1. Training Stage:

1.1. Image patching and concatenating classification, T_c and distillation token, T_d 1.2. Feeding into transformer encoder, T_E

$$T_c = T_e(T_c)$$

$$T_d = T_e(T_d)$$

1.3. Prediction from classification token:

$$P_c = \text{Classification-Head}(T_c)$$

Prediction from distillation token:

$$P_d = \text{Distillation-Head}(T_d)$$

Prediction from teacher model: P_T

1.4. Student cross-entropy loss:

$$S_{Loss} = L_{CE}(P_c, y)$$

Teacher cross-entropy loss:

$$T_{Loss} = L_{CE}(P_T, P_d)$$

Total cross-entropy loss: Average of

 S_{Loss} and T_{Loss}

$$T_l = 0.5 * (S_{Loss} + T_{Loss})$$

2. Testing Stage:

Predicted class: Average of tokens prediction

$$P_F = 0.5 * (P_c + P_d)$$

#Multi-Head External-attention Block in Transformer Encoder

1. Attention Map: Normalized patch tokens is matrix-multiplied with key memory and double normalized.

$$A_m = \text{Double} - \text{Norm}(T_{in-norm} M_k^T)$$

2. External Attention: Attention map is matrix-multiplied with value memory.

$$T_{out} = A_m M_v$$

3. Multi-Head External Attention: T_{out} for each head is concatenated.**#Teacher Model**

1. Gated Pooling:

#This operation is repeated twice

1.1. Extracted features from image, I :

$$F_r = \text{Conv} + \text{ReLU}(I)$$

1.2. Mixing proportion, $\beta = \sigma(M^T F_1)$

$$P_{gated} = \beta * \text{Max-pooling}(F_1) + (1 - \beta) * \text{Avg-pooling}(F_1)$$

2. Convolution Operation:

$$F_2 = \text{Conv} + \text{ReLU}(\text{Conv} + \text{ReLU}(P_{gated}))$$

3. G_{ap} = Global-average-pooling (F_2)4. F_c = Dense (Dense (G_{AP}))5. Classification: $P_T = \text{Soft-max}(\text{Dense}(F_c))$

to less computational effort. However, to make the model more responsive to the characteristics present in the features extracted by the Conv layer, a gated max-average pooling layer is employed. The gated pooling function provides a boost of invariance properties compared to traditional pooling which results in reducing the inductive biasing problem [40]. Initially, the input raw images are fed into the convolutional (Conv) layer and rectified linear unit (ReLU) activation function. Then, a gated pooling operation is performed by the dot product of a gating mask, M , and Conv features, x . Finally, fed it into the sigmoid function, σ to get the mixing proportion of max pooling, P_{max} and average pooling, P_{avg} to produce gated pooling, P_{gated} .

$$P_{gated}(x) = \sigma(M^T x) P_{max}(x) + (1 - \sigma(M^T x)) P_{avg}(x) \quad (11)$$

Therefore, the mixing proportion is varied depending on the characteristics of the region being pooled. In the teacher model, at first two times, the Conv and gated pooling operation are performed and then used two consecutive Conv layers followed by the RELU activation function. Later, the use of two consecutive Conv layers without utilizing the pooling layer reduces the number of parameters in CNN. Then,

instead of utilizing flatten layer, we use global average pooling which reduces each feature map to a single number by taking an average of all pixel values whether flatten layer makes the 2D vector into a 1D vector only. It also helps the model to reduce the number of parameters, hence reducing the overfitting problem. Then two dense layer is used to deeply connect to the neurons and another dense layer with a soft-max activation function is used for the teacher's prediction. The layer properties and parameters of the teacher network having altogether 11 layers are summarized in TABLE 3. The model is trained with the same Figshare and BraTS-21 datasets having an accuracy of 88.00% and 92.85%. The adequate performance of this teacher model helps to learn the local detail information as a distillation token to the transformer-based student model that is normally unable to capture and minimize total cross-entropy loss which ultimately leads to the reduction of misclassification.

G. STUDENT AND DISTILLATION LOSS

Student loss refers to cross-entropy loss calculated among true labels from the original dataset and prediction of classification token. The teacher model contributes to calculating cross-entropy distillation loss among the prediction

of the teacher model and distillation token. The total loss is measured by averaging the student loss and distillation loss. Therefore, the total loss, T_L is calculated by using equation (14) where P_T , P_c , P_d , and y denotes the prediction from the teacher model, classification token, distillation token, and true label from the original dataset respectively and L_{CE} is cross-entropy loss function.

$$T_L = 0.5 * (L_{CE}(y, P_c) + L_{CE}(P_T, P_d)) \quad (12)$$

where

$$L_{CE}(y, P_c) = - \sum_{i=1}^M y(i) \log(P_c(i)) \quad (13)$$

$$L_{CE}(P_T, P_d) = - \sum_{i=1}^M P_T(i) \log(P_d(i)) \quad (14)$$

The equation of cross entropy loss is depicted in equations (13) and (14) where M is the total number of classes in a dataset. For the Figshare dataset, $M \in 3$, and $M \in 4$ for the BraTS-21 dataset.

H. CLASSIFICATION

To know the class of the test images, the prediction from the classification token and distillation token are fed into the final classification layer. Equation (15) presents the mean prediction from both tokens to get the predicted class of the test image.

$$P_F = 0.5 * (P_c + P_d) \quad (15)$$

Here, P_F is the final prediction and P_c , P_d are predictions from classification and distillation tokens. The test image fed into the model is patched and concatenated with both tokens. Then the prediction from the classification and distillation head is averaged to get the final prediction.

TABLE 3. Layer properties of the teacher model of the proposed LCDEiT framework. (D1 and D2 indicate the Figshare and BraTS-21 datasets respectively).

Layer	Shape	Filter Number	Kernel Size	Parameter Number
Input	32x32x1	--	--	0
Conv+RELU	32x32x16	16	(3,3)	160+0
Gated Pooling	16x16x16	--	(2,2)	0
Conv+RELU	16x16x8	8	(3,3)	1160+0
Gated Pooling	8x8x8	--	(2,2)	0
Conv+RELU	8x8x128	128	(3,3)	9344+0
Conv+RELU	8x8x64	64	(3,3)	73792+0
Global Average Pooling	64	--	--	0
Dense	64	--	--	4160
Dense	32	--	--	2080
Dense +				99 +
Soft-Max	3	--	--	0 +
				(D1) (D2)

I. MODEL HYPERPARAMETERS

Hyperparameter selection is an important factor, to train the proposed model for superior results. TABLE 4 represents the optimal values of the hyperparameters of the

TABLE 4. Hyperparameter settings of proposed LCDEiT architecture.

Hyperparameter	Value
Loss function	Categorical Cross-Entropy
Number of epochs	30
Optimizer	AdamW
Batch size	512
Learning rate	0.00025
Weight decay	0.0001
Number of folds	10

proposed LCDEiT model. The model is compiled using AdamW optimizer with a learning rate of 0.00025 and fitted with a batch size of 512. Moreover, the model training is performed for 30 epochs per fold. Furthermore, the categorical cross-entropy is chosen as a loss function to compute student and distillation loss.

IV. EXPERIMENTS

For conducting training and testing of the proposed model, the Google Colab platform is used with Python version 3.7.13. The Model is implemented using Keras = 2.8.0 with TensorFlow = 2.8.2 framework. The NumPy = 1.21.6 and Scikit-learn = 1.0.2 packages have been used for image data preparation and evaluation respectively. During the training, the model occupied 3.80 GB RAM and 38.79 GB of disk space in the Colab environment. In this experiment, a random division of the dataset into ten approximately equal portions is taken and one part in sequence each time is used as the test set and the rest is used as the training set. For every fold, the model is fitted on the different training sets and evaluated on the other test sets. The network is trained using data shuffling in every iteration. Finally, the model evaluation matrices are estimated by taking an average of ten results. The summarization of steps in ten-fold cross-validation is depicted as follows:

- The dataset is divided into ten portions and each containing an equal number of images.
- For each fold, one portion is selected as a test set and the remaining are used as a training set. The selection of portions is changed in every fold.
- An average of ten results is taken to obtain the final result.

A. BENCHMARK DATASET

Two benchmark datasets are employed in this work. One is created by Cheng [33] and acquired from Nanfang Hospital and General Hospital, Tianjing Medical University, China. The database contains T1-weighted contrast-enhanced MRI images of 233 brain tumor patients with three different types such as Pituitary, Meningioma, and Glioma. And, another dataset developed by Baid et al. [34] contains multi-parametric magnetic resonance imaging (mpMRI) scans of 2,040 brain tumor patients with four different tumor classes including fluid attenuated inversion recovery (Flair), native T1-weighted (T1w), T1-weighted post-contrast (T1wce), and T2-weighted (T2w).

B. EVALUATION MATRICES

The most widely used performance indices such as accuracy, precision, recall, and F1-score are considered in this work for evaluating model performance in performing the classification task.

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \quad (16)$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \times 100\% \quad (17)$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \times 100\% \quad (18)$$

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{recall}} \times 100\% \quad (19)$$

where T_P , T_N refers to true positive and negative and F_P , F_N represents false positive and negative respectively. Performance measurement with accuracy utilizes each class in the dataset in an equal manner by taking into account overall true and false, positive and negative values which is effective for a balanced dataset. As real-life datasets may not always be balanced in class, it is efficient to widely use precision, recall, and F1-score as performance parameters. Precision and recall both focused on each class-wise performance in a model. on the contrary, F1-score is employed with averaging precision and recall thus leading to the assessment of the proposed model in terms of the F1-score widely.

V. RESULTS AND PERFORMANCE ANALYSIS

This section presents the results achieved from laborious experiments on two benchmark datasets. For ensuring an effective classification measurement, the training and testing datasets are contained in different folders.

A. QUANTITATIVE ANALYSIS

For the quantitative analysis of the proposed network, the class-wise measurement of precision, recall, and F1-score are evaluated as shown in TABLE 5. In terms of accuracy, all the classes in Figshare dataset achieve a quite similar accuracy of above 0.98. But for the BraTS-21 dataset, all the classes except the Flair class show similar accuracy of above 0.97. Hence, in terms of accuracy, the proposed model shows quite equal performance for both datasets. Due to the imbalanced dataset, the F1-score performance indices evaluation is needed as the results vary based on the number of samples in the corresponding class. In the Figshare dataset, the Meningioma class has a lower F1-score of 0.96 as the number of samples in that class is smaller among the three classes. The Glioma and Pituitary classes have a greater number of samples which causes better performance in terms of the F1-score of above 0.98 over the Meningioma class. On the other hand, for the BraTS-21 dataset, the Flair class shows the lowest F1-score of 90.38 compared to the other three classes. However, the rest of the three classes show an F1-core of above 0.94. Moreover, the confusion matrix of 10-fold cross-validation for Figshare and BraTS-21 datasets are shown in FIGURE 3. The misclassification rate is found 2.25% and 6.37% in

Figshare and BraTS-21 datasets respectively. The incorrect classification is more on the Meningioma class due to having fewer samples in this class of the Figshare dataset. For a similar reason, a greater number of samples in the Flair class of the BraTS-21 dataset is wrongly classified than the other three classes which leads to an increased misclassification rate over other classes. In addition, the area under the receiver operating characteristic curve (AUROC) per class for both Figshare and BraTS-21 datasets is depicted in FIGURE 4. In the Figshare dataset, the Glioma and Pituitary classes have a quite similar AUC value of 0.99 compared to the Meningioma class with an AUC value of 0.98. On the contrary, for the BraTS-21 dataset, the T2w and T1wce classes have a greater AUC value of 0.97 among the four classes. There is a much degradation of the AUC value of the Flair class which is observed 0.93.

B. QUALITATIVE ANALYSIS

TABLE 6 shows the model prediction performance on testing sample images. Some samples of MRI brain tumor images are tested on different architectures e.g. ResNet-50, ViT, Swin, DEiT & LCDEiT to measure how is the model prediction equal to the actual class. From the Figshare dataset, all three class samples are correctly predicted by our LCDEiT model. Whereas, the other four models are incapable to predict Meningioma class correctly. Moreover, in the case of Glioma class, the ResNet-50 and DEiT models incorrectly classify it as the Pituitary class. For BraTS-21 dataset, all the classes except T1w are correctly predicted by our LCDEiT model. Consequently, these correct predictions confirm the superiority of the proposed LCDEiT approach over the other models in classifying brain tumors into their specific classes. The dataset used in this work is imbalanced as the number of samples among classes is not equally distributed. Due to having a smaller number of samples the T1w class is predicted incorrectly.

TABLE 5. Class-wise performance measurement of the proposed LCDEiT Model.

		Classification Report				
		Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Classes
Figshare		98.98	98.07	98.60	98.34	Pituitary
		98.36	96.73	96.19	96.46	Meningioma
		98.85	98.81	98.74	98.77	Glioma
BraTS-21		97.90	96.32	95.34	95.83	T2w
		97.09	93.36	95.16	94.25	T1wce
		97.14	94.17	94.34	94.26	T1w
		95.24	90.87	89.88	90.38	Flair

C. COMPLEXITY ANALYSIS

TABLE 7 presents the computational complexity analysis of the proposed LCDEiT model over conventional DEiT in terms of multiply-accumulate (MAC) operation. The causes of quadratic complexity formulated by self-attention in the conventional DEiT can be reduced by introducing multi-head

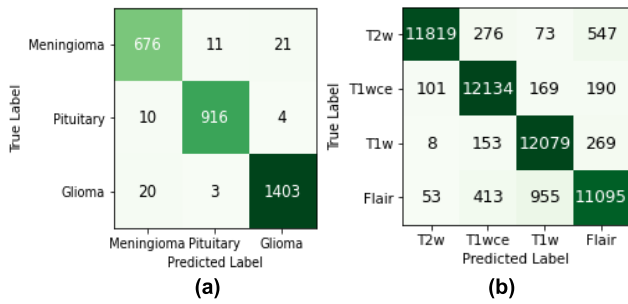


FIGURE 3. The confusion matrix of the proposed LCDEiT framework for (a) Figshare, and (b) BraTS-21 datasets.

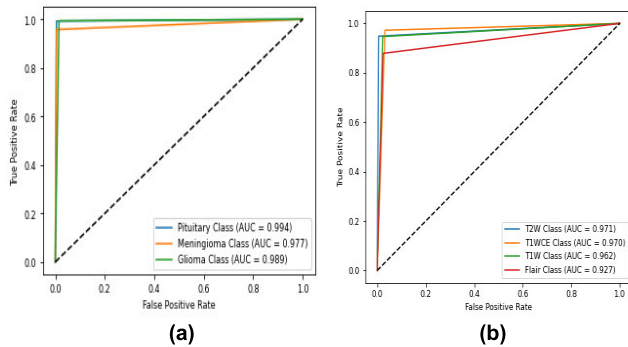


FIGURE 4. The area under the ROC curve (AUROC) of the proposed LCDEiT network for (a) Figshare, and (b) BraTS-21 dataset.

EA in the transformer encoder. For Figshare dataset, the number of parameters in MAC unit for the transformer encoder of DEiT is found 0.74M MAC but when the multi-head SA is replaced by multi-head EA, the transformer encoder needs 0.04M MAC parameters. There is around 94.30% of reduction in parameters due to the EA component. Moreover, the total trainable parameter required for the LCDEiT model is 0.43M which results in 98.68% of parameters reduction compared to the conventional DEiT model. Similarly, for BraTS-21 dataset, the trainable parameter reduction is also 98.68% for the LCDEiT model over the DEiT model. The cause of huge parameter reduction is due to using fixed valued key and value memory units in multi-head EA mechanism and a customized teacher model.

D. ABLATION STUDY

TABLE 8 depicts the impact of EA over SA by considering a gated pooled based custom CNN teacher model. The analysis is made by varying normalization techniques and values of key and value memory units. For both Figshare and BraTS-21 datasets, EA outperforms SA by improving the accuracy by 2.26% and 3.11% respectively. However, double-normalization outperforms conventional soft-max normalization for both self and external attention mechanisms. An increment of memory unit from 8 to 16 increases accuracy but a memory unit of 32 degrades the performance for both normalization techniques. Moreover, the effectiveness of the EA mechanism on image feature extraction is also

TABLE 6. Sample images with their actual class and models' predicted class. The symbols (✓) and (✗) indicate correct and incorrect predictions respectively.

Sample Image	Models	Actual Class	Predicted Class
Figshare	ResNet-50	Pituitary	Pituitary (✓)
	ViT		Pituitary (✓)
	Swin		Pituitary (✓)
	DEiT		Pituitary (✓)
	LCDEiT		Pituitary (✓)
	ResNet-50	Meningioma	Glioma (✗)
	ViT		Glioma (✗)
	Swin		Glioma (✗)
	DEiT		Glioma (✗)
	LCDEiT		Meningioma (✓)
	ResNet-50	Glioma	Pituitary (✗)
	ViT		Glioma (✓)
Swin	Glioma (✓)		
DEiT	Pituitary (✗)		
LCDEiT	Glioma (✓)		
BraTS-21	ResNet-50	T2w	T1w (✗)
	ViT		T2w (✓)
	Swin		T2w (✓)
	DEiT		T1w (✗)
	LCDEiT		T2w (✓)
	ResNet-50	T1wce	T1wce (✓)
	ViT		T1wce (✓)
	Swin		T1wce (✓)
	DEiT		T1wce (✓)
	LCDEiT		T1wce (✓)
	ResNet-50	T1w	T2w (✗)
	ViT		T2w (✗)
Swin	T2w (✗)		
DEiT	T2w (✗)		
LCDEiT	T2w (✗)		
ResNet-50	Flair	T2w (✗)	
ViT		T2w (✗)	
Swin		T2w (✗)	
DEiT		Flair (✓)	
LCDEiT		Flair (✓)	

TABLE 7. The computational complexity of LCDEiT framework in terms of multiply-accumulate (MAC) operations on two benchmark datasets.

Models	Figshare		BraTS-21	
	DEiT (MAC)	LCDEiT (MAC)	DEiT (MAC)	LCDEiT (MAC)
Transformer Encoder	740829	42208	740829	42208
Student Model	9040326	338502	9040456	338632
Teacher Model	23593859	90795	23593892	90828
Overall Architecture	32634185	429297	32634348	429460

analyzed by considering two samples of raw images from two datasets and performing both self and external attention tasks on it as shown in TABLE 9. Where SA removed all the relevant pixels in the target brain region whereas EA keeps all the relevant pixels in that area. Therefore, after performing multi-head attention this single attention map output will particularly be focused on tumor shape-based feature extraction only. In addition, TABLE 10 indicates that the proposed gated-pooled CNN achieves high accuracy by 4.51% as compared to traditional RegNetY-16GF and conventional max-pooled CNN. As experiments show the advantage of the

TABLE 8. Effectiveness of external attention mechanism over traditional self-attention with gated-pooled CNN as teacher model.

Attention	Norm	Memory Unit	Accuracy (%)	
			Figshare	BraTS-21
Self-attention	Soft-max	--	94.92	89.23
	Double-Norm	--	95.85	90.58
External-attention	Soft-max	8	97.35	91.76
		16	97.92	91.97
	32	97.81	91.85	
	Double-Norm	8	97.78	92.57
		16	98.11	93.69
	32	97.89	92.78	

TABLE 9. Impact of external attention over self-attention on the basis of image feature extraction.

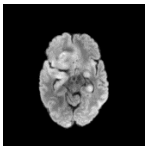
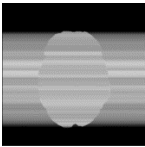
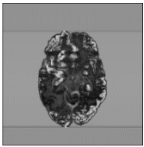
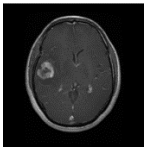
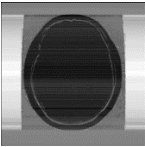
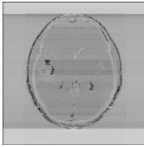
Dataset	Raw image	Self-attention	External-attention
Figshare			
BraTS-21			

TABLE 10. Performance evaluation of proposed gated-pooled based teacher model over the pre-trained model using external attention on student model.

Teacher Model	Accuracy (%)	
	Figshare	BraTS-21
RegNetY-16GF	93.60	90.67
Max-pooled CNN	95.64	91.48
Gated-pooled CNN	98.11	93.69

ability to distill knowledge from this responsive model without compromising model accuracy. This customized gated pooled CNN model is robust and generalized to compensate for the requirement of a more complex model as a teacher model.

VI. DISCUSSION

As there is a GPU memory space constrained, the image scaling has been performed with proper parameters tuning on the proposed LCDEiT model. However, concerning the practical feasibility of the proposed model such as in the medical imaging field, an assessment proposed framework by varying image size and patch size is analyzed in TABLE 11. A constant patch matrix (16×16) is utilized for running the model with a maximum of (256×256) image size. Where variation of image size from lower to larger results in deviation of accuracy values (2-3)% without parameter tuning of

TABLE 11. Impact of varying image size and patch size on the proposed LCDEiT framework (Bold values indicate with parameter tuning accuracy).

Image Size	Patch size	Patch matrix	Accuracy (%)	
			Figshare	BraTS-21
256×256	16	16×16	95.06	90.68
128×128	8	16×16	95.71	91.47
64×64	4	16×16	97.32	93.02
32×32	2	16×16	98.11	93.69

TABLE 12. Statistical assessment using P-value by Wilcoxon test of proposed LCDEiT and other models.

Model	P-value	
	Figshare	BraTS-21
ResNet-50	0.0042	0.000493
ViT	0.0040	0.000490
Swin	0.0041	0.000492
DEiT	0.0041	0.000492
LCDEiT	0.0039	0.000488

the model. If the parameter tuning will be applied, the model accuracy will be similar to the (32×32) image size. Concerning these quite lower significant changes in the results, it can be stated that the developed LCDEiT model is generalized and practically feasible where larger resolution images are preferred. For further assessment of the proposed LCDEiT model, a statistical analysis is performed based on Wilcoxon test [41] to determine the p-value as presented in TABLE 12. The TABLE implies that P-value of LCDEiT model is less than all the other models with a value of 0.039 and 0.000488 for Figshare and BraTS-21 datasets respectively.

TABLE 13. Performance comparison of different classifiers concerning test accuracy. Here, SM and TM denote the student model and teacher model respectively.

Model	Test accuracy	
	Figshare	BraTS-21
ResNet-50	92.33%	86.02%
Vision Transformer	94.33%	87.96%
DEiT (SM) + RegNetY-16GF (TM) + Self Attention	92.63%	86.49%
DEiT (SM) + Gated-Pooled CNN (TM) + Self Attention	94.92%	88.78%
DEiT (SM) + RegNetY-16GF (TM) + External Attention	93.60%	87.74%
DEiT (SM) + Gated-Pooled CNN (TM) + External Attention (Proposed)	98.11%	93.69%

In addition, the performance comparison of different classifiers concerning testing accuracy is performed as presented in TABLE 13. This table implies that the proposed model i.e. DEiT with a gated-pooled CNN teacher model and external attention increases classification accuracy, which ultimately leads to reducing misclassification. Moreover, the transformer gives much better classification accuracy than the ResNet-50 transfer learning model. The high performance of the vision transformer is limited to the large-sized dataset. DEiT acknowledges this problem and solves the dependencies of sufficient data but accuracy is compromised slightly. DEiT with a RegNetY-16GF teacher model gives

1.7% and 1.5% less accuracy than the vision transformer for both Figshare and BraTS-21 datasets respectively. The customized DEiT with a generalized and robust teacher model provides an improvement of accuracy of 2.29% than conventional DEiT for both datasets. The customized DEiT with multi-head external attention provides improved accuracy of 98.11% and 93.69% for Figshare and BraTS-21 datasets respectively.

Furthermore, the comparative analysis of the proposed LCDEiT model with the state-of-the-art model is summarized in TABLE 14 and TABLE 15. TABLE 14 presents all the existing models' comparison that uses Figshare data and BraTS-21 data utilized existing models are presented in TABLE 15. The comparison shows that the proposed LCDEiT model for both datasets provides an improvement in accuracy over another existing technique. In TABLE 14, a customized CNN is employed in [35] for the classification of Figshare data which acquires an accuracy of 95.40%. The performance is degraded by 1.2% in [10] when the genetic algorithm is employed to choose the proper parameter for the network. However, a pre-trained model called ResNet-50 with global average pooling is utilized in [36] and the accuracy is found 97.48%. Another customized CNN is used in [38] with an accuracy of 96.13%. A hybrid model of CNN along with the NADE (neural autoregressive distribution estimation) achieves 95% accuracy. However, the proposed LCDEiT provides 98.11% in Figshare which indicates superiority as compared to the other. On the contrary, in TABLE 15, a pre-trained model called EfficientNetB0 is developed in [37] for the classification of the BraTS-21 dataset with an accuracy of 55.90%. The accuracy is improved drastically by 33.20% when YOLOv5 is utilized for classification purposes. There is around 2% increment in

TABLE 14. Comparative analysis of the proposed model with the existing state-of-the-art techniques for Figshare dataset.

Model	Test accuracy Figshare Dataset
CNN [35]	95.40%
CNN & Genetic algorithm (GA) [10]	94.20%
ResNet-50 & global average pooling (GAP) [36]	97.48%
CNN [38]	96.13%
CNN & neural autoregressive distribution estimation (NADE) [39]	95.00%
LCDEiT (Proposed)	98.11%

TABLE 15. Comparative analysis of the proposed model with the existing state-of-the-art techniques for BraTS-21 dataset.

Model	Test accuracy BraTS-21
EfficientNetB0 [37]	55.90%
YOLOv5 [42]	88.00%
VGG19 [43]	90.03%
SVM [44]	84.10%
Pretrained CNN [45]	92.67%
LCDEiT (Proposed)	93.69%

accuracy when finetuned VGG19 network is used. The degradation of accuracy is observed for utilizing a machine learning algorithm named support vector machine (SVM). The model using pre-trained CNN with correlation-based selection provides 92.67% accuracy. Moreover, LCDEiT has superior accuracy 93.69% over the existing models for BraTS-21 dataset.

VII. CONCLUSION

This paper presents a teacher-student-based LCDEiT framework for categorizing tumors from brain MRIs. The framework consists of a gated-pooled CNN-based teacher model for knowledge extraction followed by image classification with an external attention-based image transformer backbone. The knowledge taken from the teacher model has compensated for the requirement of the vast dataset of vision transformers. The quadratic complexity due to self-attention in the transformer encoder is eliminated by appending external attention in the backbone transformer model that reduces complexity linearly w.r.t the number of patches. The results show that the proposed framework with the backbone of a transformer-based student model achieves the best classification performance with an F1-score of 0.978 and 0.937 for Figshare and BraTS-21 datasets respectively. This reflects the strong applicability of image transformers with a robust learner in the medical imaging field where faster computation is a crucial criterion to initiate treatment of the critical patient. In the future, the imbalance dataset handling approach such as class-wise augmentation could be implemented to overcome the issues related to a greater misclassification rate for lower sample classes. Although the proposed LCDEiT model outperformed for two distinct Figshare and BraTS-21 datasets, the experimental database could be increased further to improve the model's universality.

REFERENCES

- [1] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Phys. Med. Biol.*, vol. 58, no. 13, pp. R97–R129, Jun. 2013.
- [2] A. Hizukuri, R. Nakayama, M. Nara, M. Suzuki, and K. Namba, "Computer-aided diagnosis scheme for distinguishing between benign and malignant masses on breast DCE-MRI images using deep convolutional neural network with Bayesian optimization," *J. Digit. Imag.*, vol. 34, no. 1, pp. 116–123, Feb. 2021.
- [3] N. V. Shree and T. N. R. Kumar, "Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network," *Brain Inform.*, vol. 5, no. 1, pp. 23–30, 2018.
- [4] T. Kalaiselvi and S. T. Padmapriya, "Brain tumor diagnostic system—A deep learning application," in *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches*. Hoboken, NJ, USA: Wiley, 2021, pp. 69–90.
- [5] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel, R. Gillies, O. Gevaert, and R. Gatenby, "Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches," *Amer. J. Neuroradiol.*, vol. 39, no. 2, pp. 208–216, Feb. 2018.
- [6] V. Nardone, P. Tini, M. Biondi, L. Sebaste, E. Vanzi, G. De Otto, G. Rubino, T. Carfagno, G. Battaglia, P. Pastina, A. Cerase, L. N. Mazzoni, F. B. Buonamici, and L. Pirtoli, "Prognostic value of MR imaging texture analysis in brain non-small cell lung cancer oligo-metastases undergoing stereotactic irradiation," *Cureus*, vol. 8, no. 4, p. 584, Apr. 2016.

- [7] P. M. S. Raja and A. V. Rani, "Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach," *Biocybernetics Biomed. Eng.*, vol. 40, no. 1, pp. 440–453, Jan. 2020.
- [8] M. A. Ansari, R. Mehrotra, and R. Agrawal, "Detection and classification of brain tumor in MRI images using wavelet transform and support vector machine," *J. Interdiscipl. Math.*, vol. 23, no. 5, pp. 955–966, Jul. 2020.
- [9] P. Kaur, G. Singh, and P. Kaur, "Classification and validation of MRI brain tumor using optimised machine learning approach," in *Proc. 1st Int. Conf. Data Sci., Mach. Learn. Appl.* Cham, Switzerland: Springer, 2019, pp. 172–189.
- [10] A. K. Anaraki, M. Ayati, and F. Kazemi, "Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms," *Biocybernetics Biomed. Eng.*, vol. 39, no. 1, pp. 63–74, Jan./Mar. 2019.
- [11] S. Khawaldeh, U. Pervaiz, A. Rafiq, and R. Alkhaldeh, "Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks," *Appl. Sci.*, vol. 8, no. 1, p. 27, Dec. 2017.
- [12] Y. Yang, L.-F. Yan, X. Zhang, Y. Han, H.-Y. Nan, Y.-C. Hu, B. Hu, S.-L. Yan, J. Zhang, D.-L. Cheng, X.-W. Ge, G.-B. Cui, D. Zhao, and W. Wang, "Glioma grading on conventional MR images: A deep learning study with transfer learning," *Frontiers Neurosci.*, vol. 12, p. 804, Nov. 2018.
- [13] A. Rehman, S. Naz, M. I. Razzak, F. Akram, and M. Imran, "A deep learning-based framework for automatic brain tumors classification using transfer learning," *Circuits, Syst., Signal Process.*, vol. 39, no. 2, pp. 757–775, Feb. 2020.
- [14] T. Sadad, A. Rehman, A. Munir, T. Saba, U. Tariq, N. Ayesha, and R. Abbasi, "Brain tumor detection and multi-classification using advanced deep learning techniques," *Microsc. Res. Technique*, vol. 84, no. 6, pp. 1296–1308, 2021.
- [15] A. Behura, "The cluster analysis and feature selection: Perspective of machine learning and image processing," in *Data Analytics in Bioinformatics: A Machine Learning Perspective*. Hoboken, NJ, USA: Wiley, 2021, pp. 249–280.
- [16] L. J. Marcos-Zambrano, K. Karadzovic-Hadziabdic, T. L. Turukalo, P. Przymus, V. Trajkovic, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, and T. Klammsteiner, "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Frontiers Microbiol.*, vol. 12, p. 313, Feb. 2021.
- [17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
- [18] P. Ghosal, L. Nandanwar, S. Kanchan, A. Bhadra, J. Chakraborty, and D. Nandi, "Brain tumor classification using ResNet-101 based squeeze and excitation deep neural network," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, Feb. 2019, pp. 1–6.
- [19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
- [22] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," 2021, *arXiv:2105.02358*.
- [23] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, and M. Lucic, "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
- [24] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
- [25] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.
- [26] A. Trockman and J. Z. Kolter, "Patches are all you need?" 2022, *arXiv:2201.09792*.
- [27] G. Wang, Y. Zhao, C. Tang, C. Luo, and W. Zeng, "When shift operation meets vision transformer: An extremely simple alternative to attention mechanism," 2022, *arXiv:2201.10801*.
- [28] L. Zhang and Y. Wen, "A transformer-based framework for automatic COVID-19 diagnosis in chest CTs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 513–518.
- [29] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Inf. Fusion*, vol. 91, pp. 376–387, Mar. 2023.
- [30] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
- [31] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.
- [32] H. Wang, Y. Ji, K. Song, M. Sun, P. Lv, and T. Zhang, "ViT-P: Classification of genitourinary syndrome of menopause from OCT images based on vision transformer models," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
- [33] J. Cheng, "Brain tumor dataset," Tech. Rep., 2017, doi: [10.6084/m9.figshare.1512427.v5](https://doi.org/10.6084/m9.figshare.1512427.v5).
- [34] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, and L. M. Prevedello, "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," 2021, *arXiv:2107.02314*.
- [35] M. M. Badža and M. C. Barjaktarović, "Classification of brain tumors from MRI images using a convolutional neural network," *Appl. Sci.*, vol. 10, no. 6, p. 1999, 2020.
- [36] R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 13429–13438, Apr. 2021.
- [37] B.-H. Kim, H. Lee, K. S. Choi, J. G. Nam, C.-K. Park, S.-H. Park, J. W. Chung, and S. H. Choi, "Validation of MRI-based models to predict MGMT promoter methylation in gliomas: BraTS 2021 radiogenomics challenge," *Cancers*, vol. 14, no. 19, p. 4827, Oct. 2022.
- [38] H. H. Sultan, N. M. Salem, and W. Al-Atabany, "Multi-classification of brain tumor images using deep neural network," *IEEE Access*, vol. 7, pp. 69215–69225, 2019.
- [39] R. Hashemzahi, S. J. S. Mahdavi, M. Kheirabadi, and S. R. Kamel, "Detection of brain tumors from MRI images base on deep learning using hybrid model CNN and NADE," *Biocybernetics Biomed. Eng.*, vol. 40, no. 3, pp. 1225–1232, Jul. 2020.
- [40] C. Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 464–472.
- [41] N. Brancati, G. De Pietro, M. Frucci, and D. Riccio, "A deep learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images," *IEEE Access*, vol. 7, pp. 44709–44720, 2019.
- [42] T. Shelatkar, D. Urvashi, M. Shorfuzzaman, A. Alsufyani, and K. Lakshmana, "Diagnosis of brain tumor using light weight deep learning model with fine-tuning approach," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–9, Jul. 2022.
- [43] A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, and Z. Mehmood, "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification," *Microsc. Res. Technique*, vol. 84, no. 7, pp. 1389–1399, Jul. 2021.
- [44] P. Dequidt, P. Bourdon, B. Tremblais, C. Guillemin, B. Gianelli, C. Boutet, J.-P. Cottier, J.-N. Vallée, C. Fernandez-Maloigne, and R. Guillemin, "Exploring radiologic criteria for glioma grade classification on the BraTS dataset," *IRBM*, vol. 42, no. 6, pp. 407–414, Dec. 2021.
- [45] A. Rehman, M. A. Khan, T. Saba, Z. Mehmood, U. Tariq, and N. Ayesha, "Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture," *Microsc. Res. Technique*, vol. 84, no. 1, pp. 133–149, 2021.



GAZI JANNATUL FERDOUS is currently pursuing the B.Sc. degree in electronics and telecommunication engineering (ETE) with the Chittagong University of Engineering and Technology (CUET), Bangladesh. Her research interests include computer vision, deep learning, and data science.



KHALEDA AKHTER SATHI received the B.Sc. degree in electronics and telecommunication engineering (ETE) from the Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh, in 2019. She is currently working as a Lecturer with the Department of ETE, Chittagong University of Engineering and Technology (CUET). Her research interests include brain stimulation, computer vision, and deep learning.



MD. AZAD HOSSAIN (Member, IEEE) received the B.Sc. degree from the Department of Electrical and Electronic Engineering, Rajshahi University of Engineering and Technology (RUET), in 2004, and M.Sc. and Ph.D. degrees from Saga University, Saga, Japan, in 2010 and 2013, respectively. He is currently serving as a Professor with the Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology (CUET). His research interests include antenna for polarization switching and detection, antenna for biomedical application, brain stimulation, RF energy harvesting, and 5G antenna and MIMO antenna design.



MOHAMMED MOSHIUL HOQUE (Senior Member, IEEE) received the Ph.D. degree from the Department of Information and Computer Sciences, Saitama University, Japan, in 2012. He is currently a Distinguished Professor with the Department of Computer Science and Engineering (CSE), Chittagong University of Engineering and Technology (CUET). He is also serving as the Dean of the Faculty of Electrical and Computer Engineering (ECE), CUET, where he is the Director of the Natural Language Processing Laboratory. He has published more than 155 publications in several international journals, books, and conferences. His research interests include computer vision, human-computer interaction, and natural language processing. He served as a TPC member for several international conferences. He is a fellow of the Institute of Engineers, Bangladesh, and a Senior Member of IEEE RAS, IEEE SPS, IEEE CS, and IEEE WIE Affinity Group. He has served as the Award Coordinator, from 2016 to 2017, a Conference Coordinator, from 2017 to 2018, and a Vice-Chair Technical, from 2019 to 2021 for IEEE Bangladesh Section. He has also served as the TPC Chair for IEEE r10 HTC 2017, ECCE 2019, and ACMI 2021, the TPC Co-Chair for ICISSET 2018/22, IEEE TenSymp 2020, IEEE WIECON-ECE 2021, and ICREST 2021, and the Publication Chair for IEEE WIECON-ECE 2018/2019 and IEEE TenSymp 2020. He is also serving as the Chair of IEEE Bangladesh Section.



M. ALI AKBER DEWAN (Member, IEEE) received the B.Sc. degree in computer science and engineering from Khulna University, Bangladesh, in 2003, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2009. From 2003 to 2008, he was a Lecturer at the Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Bangladesh, where he was an Assistant Professor, in 2009. From 2009 to 2012, he was a Postdoctoral Researcher at Concordia University, Montreal, QC, Canada. From 2012 to 2014, he was a Research Associate at the École de Technologie Supérieure, Montreal. He is currently an Associate Professor with the School of Computing and Information Systems, Athabasca University, Athabasca, AB, Canada. He has published more than 70 articles in high impact journals and conference proceedings. His research interests include artificial intelligence, affective computing, computer vision, data mining, information visualization, machine learning, biometric recognition, medical image analysis, and health informatics. He has served as an editorial board member, a chair/co-chair, and a TPC member for several prestigious journals and conferences. He received the Dean's Award and the Excellent Research Achievement Award for his excellent academic performance and research achievements during his Ph.D. studies in South Korea.

...