Wiley
International Journal of Intelligent Systems
Volume 2025, Article ID 1444673, 15 pages
https://doi.org/10.1155/int/1444673

WILEY

Research Article

X-SCSANet: Explainable Stack Convolutional Self-Attention Network for Brain Tumor Classification

Rahad Khan and Rafiqul Islam

Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur 1707, Bangladesh

Correspondence should be addressed to Rafiqul Islam; rafiqul.islam@duet.ac.bd

Received 4 November 2024; Accepted 8 March 2025

Academic Editor: Eugenio Vocaturo

Brain tumors are devastating and shorten the patient's life. It has an impact on the physical, psychological, and financial well-being of both patients and family members. Early diagnosis and treatment can reduce patients' chances of survival. Detecting and diagnosing brain cancers using MRI scans is time-consuming and requires expertise in that domain. Nowadays, instead of traditional approaches to brain tumor analysis, several deep learning models are used to assist professionals and mitigate time. This paper introduces a stack convolutional self-attention network that extracts important local and global features from a freely available MRI scan dataset. Since the medical domain is one of the most sensitive fields, end-users should put their trust in the deep learning model before automating tumor classification. Therefore, the Grad-CAM method has been updated to better explain the model's output. Combining local and global features improves brain tumor classification performance, with the suggested model reaching an accuracy of 96.44% on the relevant dataset. The proposed model's precision, specificity, sensitivity, and F1-score are reported as 96.5%, 98.83%, 96.44%, and 96.4%, respectively. Furthermore, the layers' insights are examined to acquire a deeper knowledge of the decision-making process.

Keywords: brain tumor classification; convolutional neural network; explainable AI; MRI; self-attention

1. Introduction

The medical domain is one of the research fields where the world's most funds are invested to make human beings well, improve treatment and diagnosis, and discover medicines for diseases. Brain tumors are one of the severe diseases and a patient with a brain tumor has a low survival rate. The rapid production of abnormal cells in the brain is called a tumor. Tumors affect a patient's lifetime physically as well as psychologically [1]. Magnetic resonance imaging (MRI) is a noninvasive anatomical modality that uses magnets and radio waves to capture information about bones, muscles, and tissues. One of the major applications of MRI is the detection and diagnosis of brain tumors [2]. Although there are several forms of brain tumors, we will focus on three: glioma, meningioma, and pituitary. However, these tumors can broadly be categorized as malignant and benign (nonmalignant). While benign tumors pose a lower risk, malignant tumors are life-threatening. Among these three types, gliomas are the most aggressive primary brain tumor [3]. Originating in the brain or spinal cord, gliomas have a rapid growth rate and are often malignant, posing a significant health risk. Meningiomas and pituitary tumors, on the other hand, are typically benign and may originate in the skull or spine [4]. Although all tumors are not threatening, detection of the type of tumor is crucial for further diagnosis. Traditionally, radiologists manually detect brain tumors from MRI images which is time-consuming and can be erroneous owing to unwanted artifacts in the images [5]. Deep learning, a subfield of machine learning, excels at analyzing unstructured data such as images and has emerged as a powerful tool for automated brain tumor detection, offering faster and potentially more accurate results.

An image is a spatial representation of pixels where each pixel contains an intensity value. Humans are experts at understanding the context within an image, readily detecting and recognizing objects from the image. During this process, our visual system strategically shifts its focus, sometimes examining the entire image and other times zooming in on specific regions to extract crucial information. Features that capture information about the whole image, such as color distribution or overall shape, are called global features. In contrast, features that focus on smaller regions, such as edges, corners, or textures, are known as local features. An image has a combination of both local and global features [6]. Convolutional neural networks (CNNs) excel at capturing local features due to their core operation, the convolution. During convolution, a small filter scans the image, extracting information from localized regions. While self-attention mechanisms focus on establishing relationships between individual pixels and the entire image, allowing the capture of global features [7]. Integrating multiple machine learning models to achieve a common goal is known as an ensemble technique. Ensemble approaches include bagging, boosting, stacking, and blending. Stacking, in particular, is an ensemble technique where the input is passed through multiple models, and the results of those models are concatenated and fed into another model [8]. This ensembling process enhances the performance of the machine learning model.

Nowadays, machine learning is applied in various domains and the performance of these models is undoubtedly excellent. However, the internal decision-making processes are a serious concern to the researchers since it seems to be a black box. In addition, this lack of transparency raises doubts about the reliability and trustworthiness of these models in real-world applications. To resolve this issue, researchers introduced a new field called explainable artificial intelligence (XAI), which integrates several algorithms to interpret model behaviors and provide explanations behind the model predictions [9–11]. As we mentioned earlier, XAI has a variety of algorithms depending on a couple of factors such as model-specific or model-agnostic, local or global explanations, and so on. Among some popular XAI algorithms, this work utilizes the gradient-weighted class activation mapping (Grad-CAM) algorithm to generate a saliency map for the proposed model. The main benefit of the method is that it will provide valuable insights into the inner workings of the model and enhance the understanding of its decision-making processes [12].

The subsequent sections of this paper are organized as follows. Section 2 presents various research efforts by different researchers in the field of detecting tumors from brain MRI images. A comprehensive explanation of the proposed stack convolutional self-attention network (SCSAN) model is provided in detail in Section 3. The performance of the proposed model, evaluated using several metrics, is discussed in Sections 4 and 5, concluding our work.

2. Related Works

The field of medical imaging segmenting [13, 14], detecting, and classifying [15, 16] brain tumors using a machine learning algorithm is in demand. CNNs have revolutionized the field of computer vision, offering a powerful tool for image-based tasks. When dealing with image recognition problems that require learning and extracting important features, CNNs are often the first machine learning model that comes to mind. The authors in [17] proposed a parallel deep CNN (PDCNN) with two distinct CNN architectures. Each CNN architecture employs kernels of different sizes to capture local and global features. Three publicly available datasets and dimensions of input images 32×32 are used to train PDCNN and the accuracy they achieved is 97.33%, 97.60%, and 98.12%, respectively. The authors introduced a similar framework in [18] to demonstrate the efficacy of CNN and autoencoder-based algorithms for classifying multiclass cancers using MRI–based neuroimaging. The learning procedure took into account both weighted grayscale and blue contrast MR images in order to achieve optimum diagnostic accuracy. However, the model's main problem was the limited usage of advanced architecture or upgraded datasets, which could diminish the algorithm's generalization potential. The authors [19] proposed a novel architecture, DeepTumorNet, which combines GoogLeNet with additional layers, including clipped leaky rectified linear unit (ReLU) and group convolution, for brain tumor classification. However, DeepTumorNet has some potential limitations. It could be computationally expensive, and its performance may be susceptible to overfitting, as the authors do not mention the use of regularization techniques to mitigate this risk. Focusing on automating brain tumor detection, the authors [20] employed a deeper CNN architecture with small kernels to reduce model complexity while maintaining effectiveness. The CNN was used for binary tumor classification, achieving an accuracy of 97.5%. The work [21] proposed an improved CNN model for brain tumor classification, aiming to be able to be applied in an IoT healthcare environment. Two publicly available datasets from Kaggle were utilized, one for training and the other for cross-validation. Due to the limited number of MRI images in each dataset, data augmentation techniques were employed to expand the training data. The augmented images were then passed through several pretrained models (VGG16, ResNet50, GoogleNet, MobileNet, etc.) to extract features. These extracted feature maps were subsequently fed into the proposed CNN model to classify the input image as either meningioma, glioma, or pituitary tumor. Several studies [22–27] have explored the transfer learning for brain tumor analysis by utilizing various pretrained models such as AlexNet, VGG16, VGG19, ResNet18, ResNet50, ResNet101, DenseNet121, InceptionV2, InceptionV3, EfficientNet-B0–B7, SENet, and Xception to extract feature maps for classification, detection, and performance comparison. It was observed that these pretrained models can efficiently extract prominent features in less time. Although the accuracy of these models is excellent, their weights are based on the ImageNet [28] dataset, which contains images of real-world objects. The feature extraction for the ImageNet dataset is significantly different from that for medical images.

Nowadays, ensemble learning is more robust and powerful, offering higher performance compared to a single

machine learning model. To identify a MRI scan as malignant or benign, the CNN is utilized as a baseline for feature extraction, and the extracted features are further processed by the decision tree (DT) and radial basis function (RBF) [29]. Similarly, a CNN model serves as a baseline, and instead of DT and RBF as classifiers, a support vector machine (SVM) predicts whether an input image is cancerous or benign with 98.495% accuracy [30]. With the CNN architecture, we can only capture the local features of images, whereas an image comprises both local and global features. To capture these global features, a transformer with self-attention is used instead of a CNN. Initially, transformers were proposed to handle long-range dependencies in text sequences. The author [31] first proposed a transformer architecture capable of processing images. Now, transformers are also used in computer vision tasks, particularly in the medical domain [32].

Following extensive research, this work focuses on the usage of a stack CNN model with a self-attention layer for brain tumor classification. Furthermore, the Grad-CAM–based XAI algorithm analyzes model layers to better comprehend the decision-making process. Overall, the significant contributions of the work are highlighted in the field of brain tumor classification as follows:

- Proposed a novel ensemble-based deep learning model called the SCSAN, which effectively handles both local and global features.
- Comparison of the SCSAN model with other recent deep learning models in terms of parameters, accuracy, precision, and recall.
- Assurance of high performance of the SCSAN model, along with an interpretation of its internal decision-making processes by visualizing the saliency map using customized Grad-CAM.

3. Methodology

Early initiation of treatment for a patient with a brain tumor may minimize the associated risks. However, the conventional method for detecting and classifying malignancies in brain MRI scans is time-consuming and error prone. To address this challenge, we propose the utilization of deep learning algorithms, specifically a combination of CNN and self-attention. The pipeline of the proposed model is shown in Figure 1. The quality and diversity of the dataset used to train any deep learning system have a major impact on its performance.

3.1. Data Preprocessing. The MRI images in the dataset were derived from various sources, leading to differences in their dimensions. To ensure that the data were consistent, all images were resized to $128 \times 128 \times 3$ pixels through the use of a bilinear interpolation algorithm. This algorithm calculates the pixel values of the destination image by taking the weighted average of the neighboring 4×4 pixels of the source image. Since the images were represented in 8-bit grayscale and had pixel values ranging from 0 to 255, normalization was applied to scale the values between 0 and 1.

$$X = \frac{X - X \cdot \min()}{X \cdot \max()}, \tag{1}$$

where X represents an image and min () and max () are functions that return the minimum and maximum values of a NumPy array, respectively.

3.2. Model Description. The proposed SCSAN harmoniously integrates the advantages of two distinct deep learning algorithms: CNN and self-attention. CNN employs local receptive fields across an input image to extract a feature map known as local features. Conversely, self-attention identifies relationships between individual pixels and all others without relying on local context. The features extracted by self-attention are referred to as global features. An image is a spatial representation of local and global features that are helpful for image localization [33], segmentation [34], detection [15], recognition [35], and so on using deep learning approaches. The proposed model addresses both local and global features by CNN and self-attention, as illustrated in Figure 2. The input images are fed into the CNN and SA blocks simultaneously. Each CNN and SA block extracts the corresponding features. Extracted features are then fed into two independent, fully connected networks. The final outputs from these networks are stacked and passed through another fully connected network, which has four neurons that correspond to four classes of tumors.

3.2.1. CNN Block. The main building block of CNN that makes it more popular than other deep learning algorithms is convolution operation. It automates feature extraction through a process of sliding a filter across the input image and performing elementwise multiplication, capturing specific patterns.

$$\mathrm{con}v(X) = \sum_{j=1}^{n} B_i + X_j * K_{i,j}, \tag{2}$$

$$\mathrm{bn}(x) = \frac{x - E[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta, \tag{3}$$

$$\mathrm{ReLU}(x) = \max(0, x), \tag{4}$$

$$\tan h(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{5}$$

The depth of the filter typically matches the depth of the input channel. The same filter K convolves $(*)$ the input X along all channels j of the input. This process is repeated for multiple filters i. Finally, a single bias B_i is added to the output of each filter convolution to create feature maps con$v$. Building on the concept of using CNNs to extract local features from images, this paper proposes a three-layer CNN architecture, as illustrated in Figure 3. Initially, we employ 16 filters of size 5×5 with a stride of 2, allowing for features to

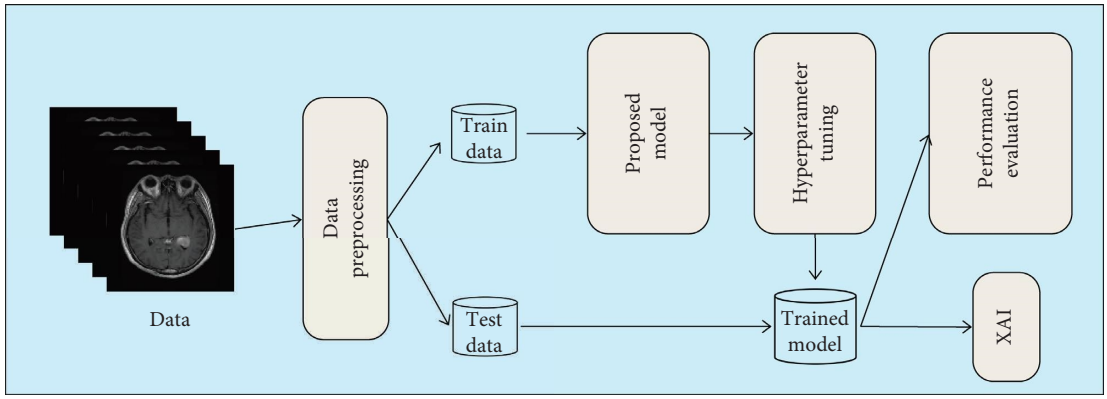

Figure 1: The pipeline for the proposed model.

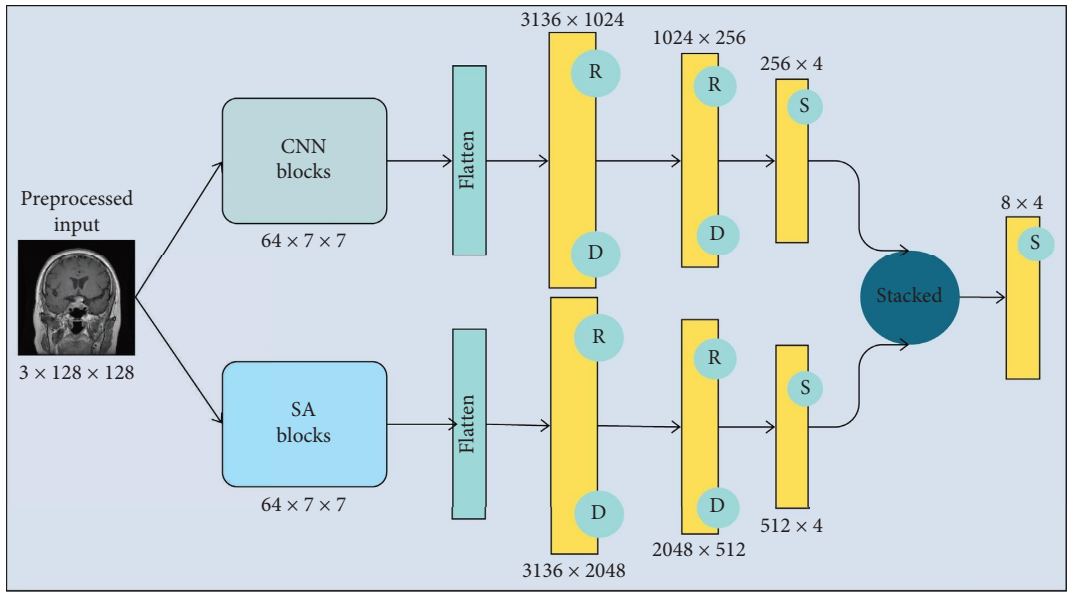

Figure 2: Proposed stack convolutional self-attention network.

overlap and capture richer details. This is followed by batch normalization and an activation function. Batch normalization (3), a technique that accelerates training towards local or global optima by stabilizing the learning process, is applied after each convolution. To introduce nonlinearity, crucial for complex pattern recognition, we employ two activation functions in different layers of the convolution block: ReLU (4) and hyperbolic tangent (Tan$h$) (5). The resulting feature maps, sized $16 \times 62 \times 62$, undergo downsampling through max pooling and proceed through another convolution operation to get the feature maps $F^{(1)}$.

$$F^{(1)} = \mathrm{ReLU}(\mathrm{bn}(\mathrm{conv}(\mathrm{max\_pool}(\mathrm{ReLU}\ (\mathrm{bn}(\mathrm{conv}(X))))))). \tag{6}$$

The output $F^{(1)}$ of the initial convolutional layer is fed into three parallel convolutional block layers. Each convolution block employs filters of different sizes and strides, allowing them to capture distinct features from the same input feature map [36]. The outputs of these three blocks are concatenated depthwise $[F^{(2_1)}, F^{(2_2)}, F^{(2_3)}]$, and 3×3 average pooling is applied for downsampling. To establish a residual connection, the output $F^{(1)}$ of the first convolutional layer is added to the result $F^{(2)}$ of average pooling. This sum $F^{(12)}$ is then input to the final convolutional layer, ensuring that the model retains important features from the previous layer [37].

$$F^{(12)} = F^{(1)} + F^{(2)}. \tag{7}$$

Following the final convolution layer, the output feature map (F_{conv}) is reshaped into a one-dimensional vector through a flattening operation. This vector is then fed into dense layers, as depicted in Figure 2.

3.2.2. Self-Attention Block. The proposed SCSAN model includes another essential component called the self-attention block. Self-attention [38] was initially designed for handling sequences in natural language processing (NLP), where it serves as the primary building block of

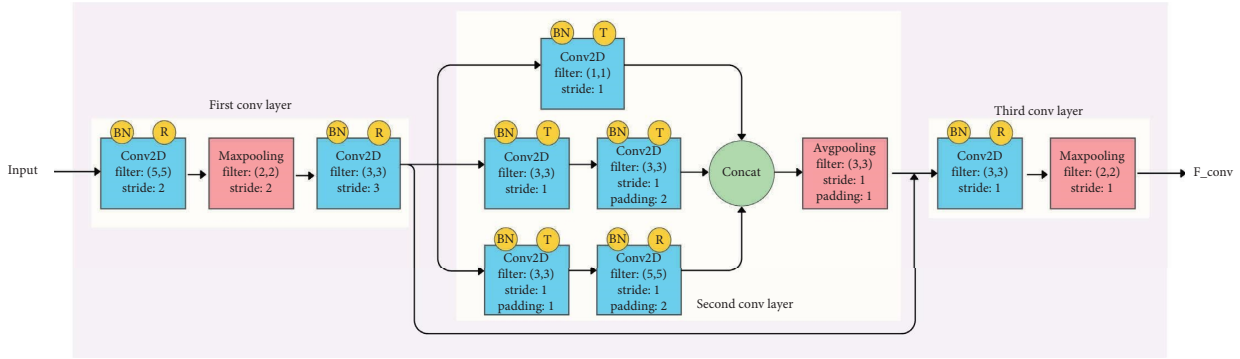

FIGURE 3: Convolution part of the proposed SCSAN model. To get F_{conv} from input, three-layer convolution is used.

a transformer. Interestingly, sequences are not limited to text. Images can also be viewed as containing sequences, particularly at the pixel level. The motivation behind choosing a raw self-attention block instead of any variants [39, 40] of a vision transformer is to make our SCSAN model simple and computationally efficient.

Since the self-attention model excels at handling sequential data, the image is converted into a sequence of tokens using patch embedding. By dividing the image into 4×4 patches, a total of 1024 smaller square tokens is created. These tokens are then processed in parallel by the self-attention mechanism. The position of each token plays a crucial role in understanding the image's structure, and we employ 1024 learnable parameters called positional embeddings. In addition, a single class embedding vector is prepended to the sequence at position zero to represent the overall image class. The self-attention block of the SCSAN model is detailed in Figure 4. The values of both positional and class embeddings are learned during the training phase. This allows the model to capture the spatial relationships within the image and the overall class information effectively.

$$X = \rho(I), \tag{8}$$

where ρ represents the patch embedding followed by positional and class embedding. Within a training batch, we consider a set of B input sequences, each containing L elements and having an embedding dimension of D. These sequences are formed by combining positional and class embedding data. The resulting combined data, denoted as $X \epsilon R^{B\times L\times D}$, is subsequently fed into the multihead self-attention (MHSA) layer. Unlike single-head self-attention, which focuses on capturing a singular type of relationship within sequences, MHSA leverages H-independent attention heads in this case, $H = 4$. This architecture allows the model to learn and attend to diverse and intricate relationships between elements within the sequence.

$$A^{(1)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \tag{9}$$

$$X = \text{ReLU}\left(\text{LN}\left(X + A^{(1)}\right)\right), \tag{10}$$

$$F\_\text{attn} = \text{LN}\left(X + A^{(2)}\right). \tag{11}$$

The core operation of MHSA is to calculate the attention scores of each pixel relative to the others using three matrices: query (Q), key (K), and value (V). Each pixel in an image determines how much attention it gives to the other pixels in the same image, which characterizes the process as self-attention. The query $Q \epsilon R^{L\times H\times H_{\text{dim}}}$ seeks information from the keys $K \epsilon R^{L\times H\times H_{\text{dim}}}$ within the sequence. The degree of similarity between a query and a key determines the relative importance of the corresponding value $V \epsilon R^{L\times H\times H_{\text{dim}}}$. Finally, the values V, weighted by these importance scores, contribute to the final output $A^{(1)}$ of the MHSA layer. To achieve a more generalized and stable learning process, the residual connection is employed. The input sequence X is elementwise added with the output $A^{(1)}$ from the first MHSA layer. This combined result is then normalized using a layer normalization technique. Similarly, the final output F_attn of the entire attention block is obtained by adding the normalized sum of the X and the output $A^{(2)}$ from the second MHSA layer.

3.2.3. Stacking. The outputs $y_{\text{conv}} \epsilon R^{B\times C}$ and $y_{\text{attn}} \epsilon R^{B\times C}$ contain the scores for different tumor classes for the input image, based on local features captured by the convolution approach and global features captured by the self-attention approach, respectively. The proposed SCSAN model then stacked y_{conv} and y_{attn} to produce $y_{\text{can}} \epsilon R^{B\times 8}$. Finally, y_{conv} is passed through another fully connected layer as depicted in Figure 2 to calculate the class probabilities for the tumor types: glioma, meningioma, pituitary, and no-tumor.

3.3. Explainable AI. In the field of medical imaging, particularly when identifying and categorizing brain tumors from MRI scans, the final decision-making role must lie with a domain expert, as patient lives are at stake. This holds true across the entire medical domain, where any incorrect diagnosis or treatment plan can have serious consequences. However, the age of artificial intelligence is driving a push for automation in healthcare as well. Researchers are actively developing deep learning models for various medical tasks. While achieving high accuracy is crucial, it is not the sole indicator of a model's suitability for real-world medical applications. The lack of transparency in a model's internal decision-making process remains

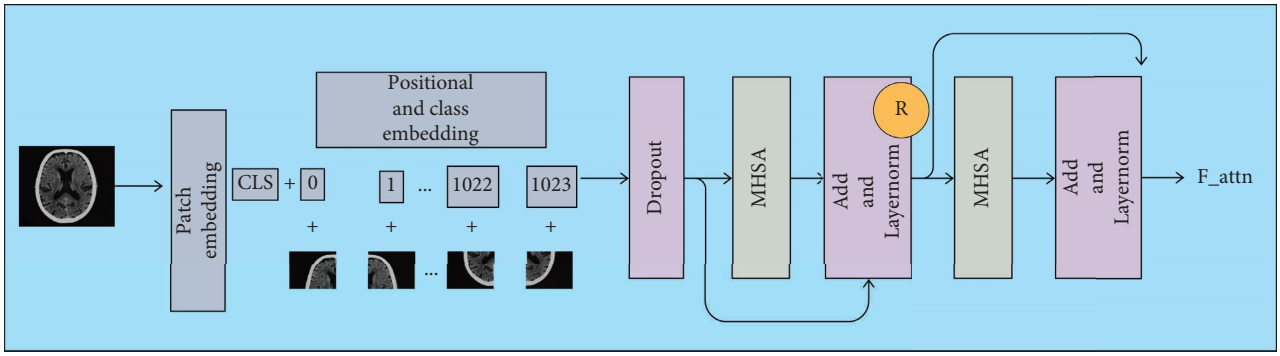

Figure 4: The self-attention block of the SCSAN model. The multihead self-attention (MHSA) block operates on the input features while maintaining their original dimensionality.

a significant hurdle. To ensure the reliability of the SCSAN model beyond just high accuracy, we need to understand its internal behavior. As our model incorporates a CNN component, visualizing the feature maps at different layers can provide insights into what features the model is detecting within the MRI scans. Figure 5 displays the feature maps for Layers 0 and 1 of the CNN portion of our proposed SCSAN model. While visualizing kernels and feature maps from the convolution operation can offer some clues about the patterns a model is looking for, similar to what an expert might do, the Grad-CAM approach provides a deeper level of understanding by revealing the specific image regions most crucial for the model's predictions. This transparency is pivotal for building trust in the model, as it allows us to move the model from a "black box" to a "glass box" approach. The process of determining a saliency map for an input image is illustrated in Figure 6. Our proposed SCSAN model incorporates both CNNs and a self-attention mechanism. To accommodate this architecture, we introduce a slight modification to the traditional Grad-CAM approach for generating heatmaps detailed in Algorithm 1. Since Grad-CAM is better suited for CNN–based models [41], we concatenate the feature maps obtained from both the CNN and self-attention parts. This concatenated feature map is then passed through a final rectified convolution layer to a generate feature map A_{ij}^{k}. Subsequently, fully connected networks are employed. Here, we focus on Class c of the input image. We calculate $(\delta y^{c}/\delta A_{ij}^{k})$ the gradients of the output score for Class c. These gradients act as weights W_{m}^{c}, which are multiplied elementwise with the corresponding elements in the feature map A_{ij}^{k}. The resulting weighted sum is then passed through a ReLU activation function, as negative values might indicate activations for other classes.

Algorithm 1 utilizes two functions get_activations and sort. The get_activations function takes a specific layer within the model as input. This layer must perform a convolution operation followed by a ReLU nonlinear activation function. The function then returns the feature map values obtained when the input image x is passed forward through the entire model. The sort function, on the other hand, takes the class scores predicted by the model and arranges them in a descending order. It returns a list of indices, where each index corresponds to a class label, sorted based on the predicted class scores from highest to lowest.

4. Results

Large datasets, while promoting performance enhancement, can also significantly increase the computational demands of deep learning models. Training a CNN and a self-attention model in parallel on a dataset [42] containing three tumor types (glioma, meningioma, and pituitary) and one nontumor type can be computationally expensive. This can lead to a slower overall training process, especially when GPU acceleration is unavailable. For training and computations, we leveraged the capabilities of a Kaggle P100 GPU with 16 GB of memory and 29 GB of RAM. The proposed SCSAN model was implemented using the PyTorch framework. We aimed to achieve high model performance while simultaneously reducing the number of learnable parameters. This resulted in a model with approximately 11.33 million parameters, which can be trained in roughly 30 min.

4.1. Dataset. Medical imaging datasets are typically less readily available compared to other fields, primarily due to concerns related to patient confidentiality, high costs, and various medical ethical considerations. The use of a supervised deep learning algorithm requires a substantial dataset with accurate annotations [33]. In this study, we used publicly available MRI datasets for brain tumors from Kaggle. The dataset [42] comprises four classes: one nontumor class (2000 samples) and three tumor classes: glioma (1621 samples), meningioma (1645 samples), and pituitary (1757 samples). We addressed the multiclassification problem of determining whether a MRI image belongs to either the nontumor class or one of the three tumor classes using the proposed SCSAN model. Figure 7 displays MRI images representing various classes of the dataset.

4.2. Hyperparameter Tuning. The performance of deep learning models is highly sensitive to hyperparameters, which are predefined parameters set by the researcher. Selecting optimal hyperparameters can significantly improve model performance. In this paper, we used GridSearchCV,

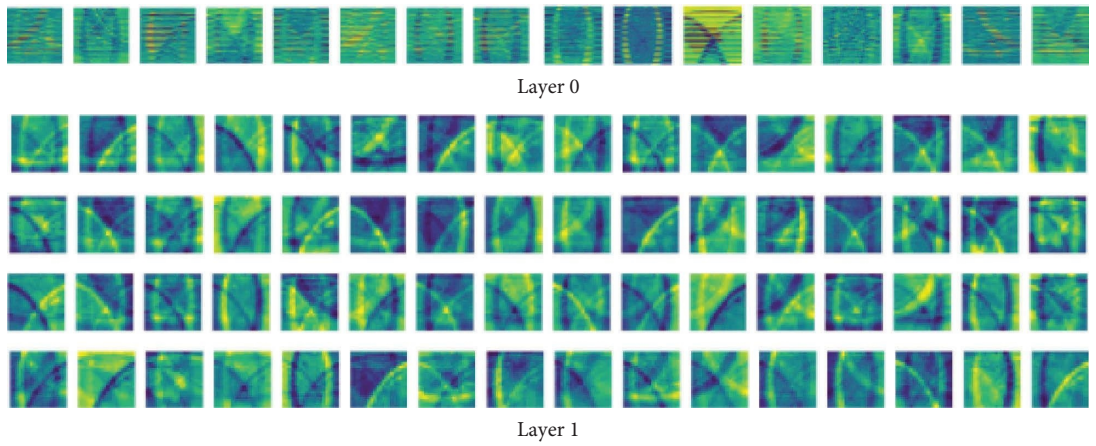

FIGURE 5: The initial layers of a convolutional operation identify basic patterns, while the deeper layers detect more complex patterns as their number increases.

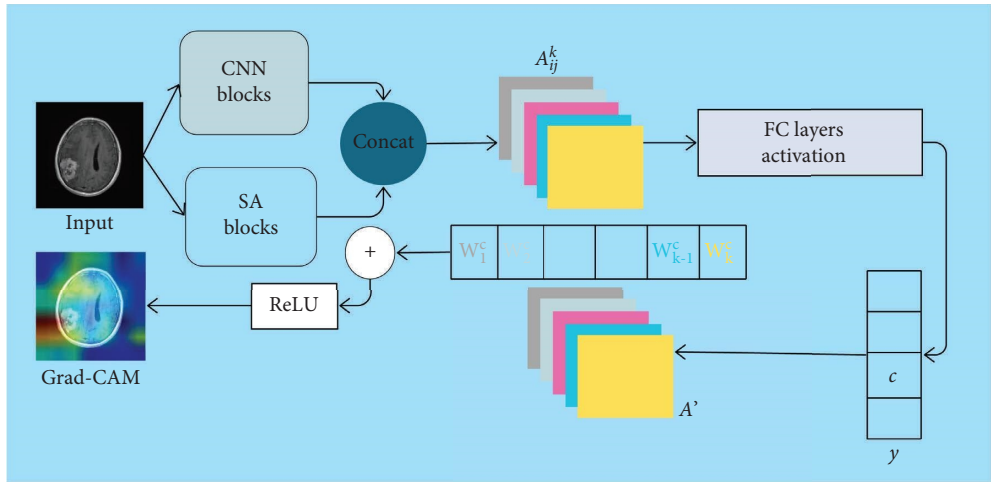

FIGURE 6: The process of obtaining a Grad-CAM heatmap for the SCSAN with an input image.

Require: model: Trained SCSAN model
layer: Target layer
x: Input image
Ensure: H: Heatmap for the input image

1. out $\longleftarrow$ model(x)
2. out $\longleftarrow$ softmax(out)
3. $A_{ij}^k \longleftarrow$ get_activations(layer)
4. $H \longleftarrow$ []
5. **for** idx in sort(out) **do**
6. y^c = out[idx]
7. $W_k^c = (1/Z)\sum_i \sum_j (\delta y^c / \delta A_{ij}^k)$
8. $L_{\text{Grad-CAM}}^c = \sum_{m=1}^k W_m^c A_{ij}^k$
9. $h = \text{ReLU}(L_{\text{Grad-CAM}}^c)$
10. $h = (h - h \cdot \min())/h \cdot \max())$
11. $H \cdot \text{append}(h)$
12. **end for**
13. **return** $H = 0$

ALGORITHM 1: Generate a heatmap for the SCSAN model by adapting the Grad-CAM method.

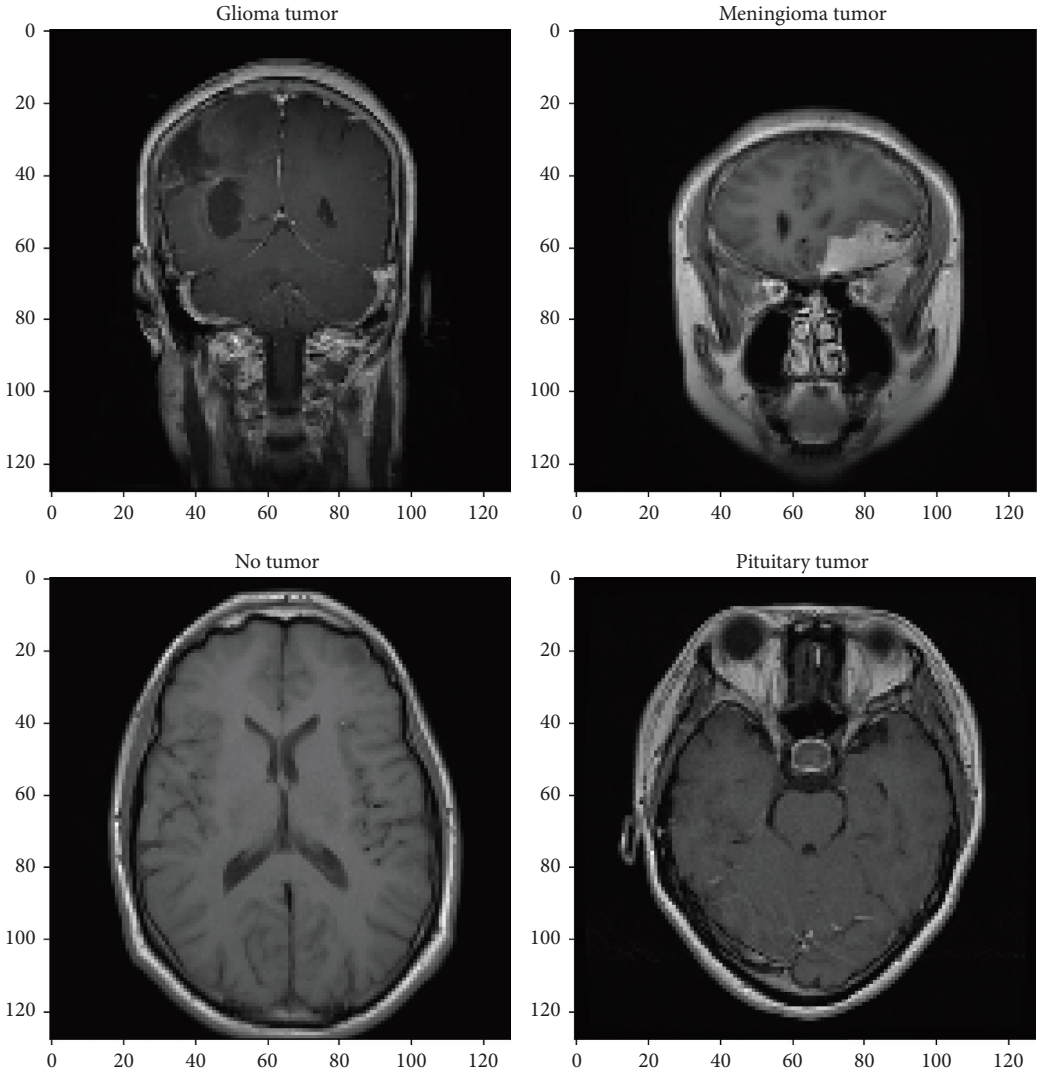

Figure 7: Images of different classes from the dataset. Glioma, meningioma, and pituitary are treated as "tumor" class while the remaining one as "no-tumor."

a hyperparameter tuning technique, to optimize the hyperparameters of the SCSAN model. The dataset was split into training 80% and testing 20% sets for model evaluation. The Adam optimizer was used with a learning rate of $1e-3$ to train the model for 100 epochs. To prevent overfitting, we employed L2 regularization and a learning rate scheduler to guide the model towards the optimal solution. Since we are addressing a multiclassification problem, the cross-entropy loss function was utilized to calculate the prediction loss of the SCSAN model. Notably, during training, the model with the highest validation accuracy was tracked and used for final performance evaluation.

4.3. Performance Evaluation. The training process yielded the best-performing SCSAN model at Epoch 71, achieving a test accuracy of 96.44% out of 100 epochs. This highlights the effectiveness of the training strategy. As expected, the test loss of the SCSAN model exhibited a significant decrease that mirrored the reduction in training loss. However, after reaching Epoch 50, the rate of loss decrease slowed down. Conversely, the accuracies for both training and test data steadily increased throughout the training process, as illustrated in Figure 8.

In the domain of classification tasks, the confusion matrix serves as a valuable tool for evaluating the performance of machine learning models. It provides insights into how effectively the model differentiates between actual class labels for unseen test images. Figure 9 depicts the confusion matrix generated for the SCSAN model during the evaluation phase. This matrix not only reveals classification patterns but also facilitates the calculation of various performance metrics such as individual class accuracy, sensitivity, precision, and so on. As illustrated in Figure 9, the SCSAN model misclassified 6 glioma tumor images as meningioma. This misclassification is concerning because glioma tumors are often malignant and accurate prediction is crucial to avoid potential risks associated with misdiagnosis for patient health. The underlying reason for these misclassifications lies in the occasional visual similarity

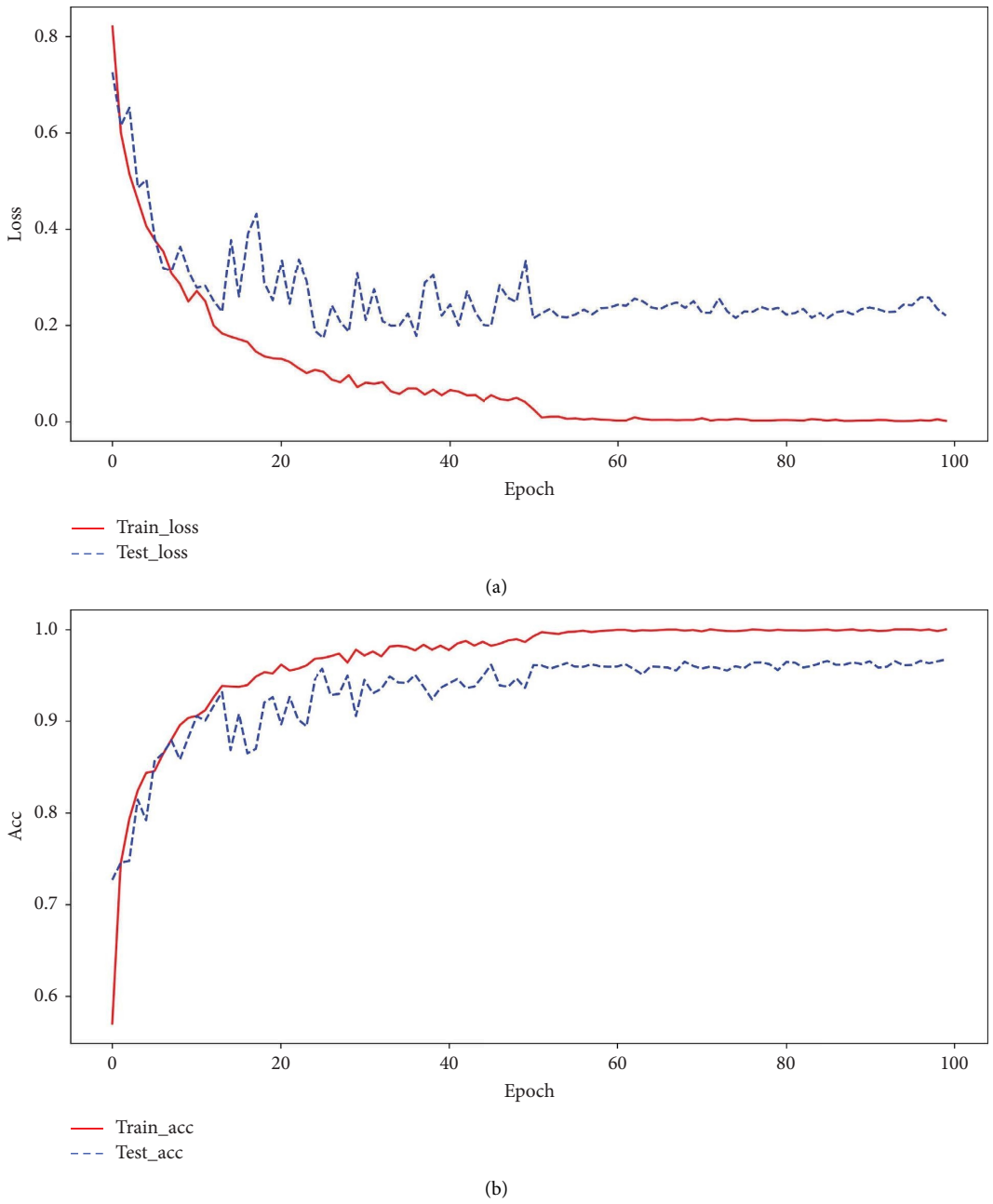

Figure 8: The solid line represents the training loss/accuracy, while the dashed line represents the test loss/accuracy. Losses and accuracy of our proposed model evolve as epochs progress during the training period. (a) Both the patterns of train and test loss exhibit a decay as epochs progress, although there is a slight instability in the test loss. (b) Training accuracy has stabilized at 99.95% after nearly 60 epochs.

between glioma and meningioma on MRI scans, especially for lower-grade tumors. The misclassification of glioma tumors as meningioma can also impact the precision and recall metrics specifically for these two classes, as illustrated in Figure 10(a). Although the receiver operating characteristic (ROC) curve is traditionally used for binary classification problems, it can be utilized for multiclass problems by considering one class as positive and the remaining classes as negative. In our case, the microaverage and macroaverage metrics for the SCSAN model are both 0.98, indicating strong overall performance. Figure 10(b) depicts the ROC curves for the proposed model. As expected, the area under the curve (AUC) for Class 3 (pituitary tumor) is the highest, mirroring the strong performance observed in the other evaluation metrics.

4.4. Performance Comparison. Brain tumors are a devastating disease, often leading to fatalities due to delayed diagnosis. Recognizing the urgency for early detection, researchers have actively explored various machine learning, particularly deep learning, algorithms to predict

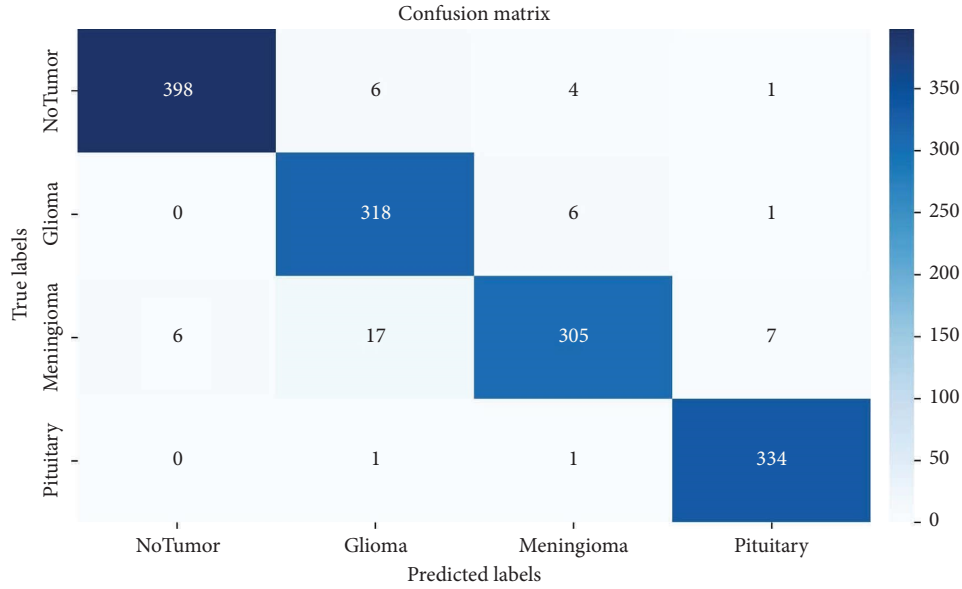

FIGURE 9: Confusion matrix of the proposed SCSAN model. Our model demonstrated strong classification capabilities in the pituitary tumor class. It correctly classified 334 out of 336 pituitary images.

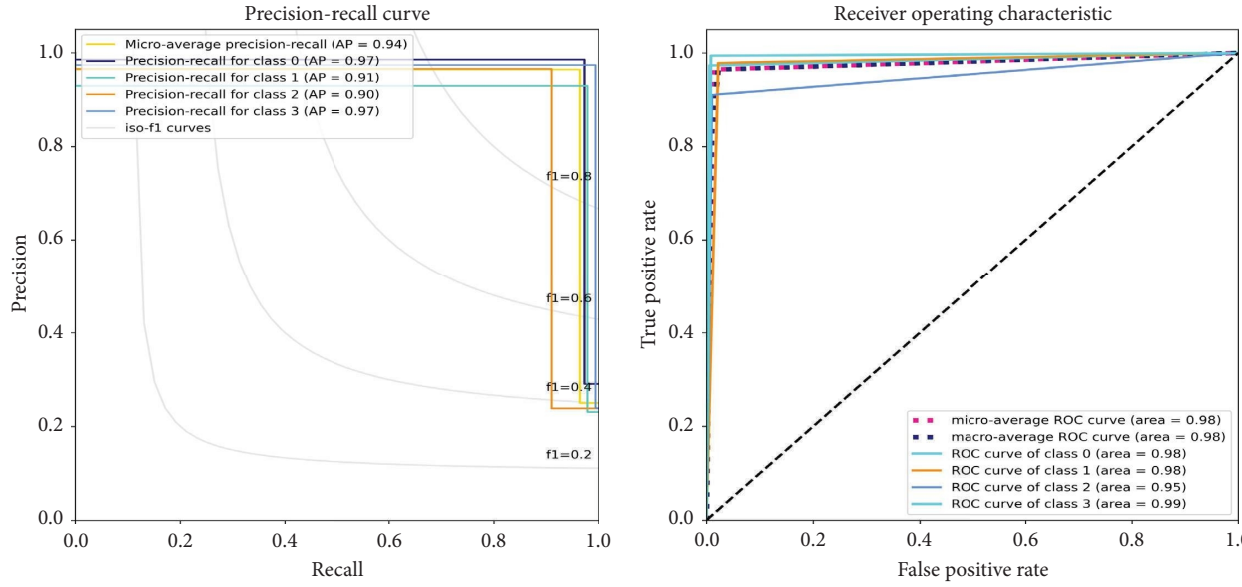

FIGURE 10: Performance analysis using recall and ROC curve. (a) The x-axis represents recall and the y-axis represents precision for the classes. A higher average precision (AP) for a class indicates that the model effectively distinguished that class from other classes. (b) The receiver operating characteristics curve for the SCSAN model. The x-axis represents the false positive rate (FPR) and the y-axis represents the true positive rate (TPR) for the classes.

the presence of tumors, especially malignant ones, in brain MRI scans. To evaluate the effectiveness of our proposed SCSAN model, we compare its performance with established CNN architectures such as VGG16, ResNet50, and EfficientNet-B7.

For our evaluation, we utilized a dataset of 7023 brain MRI scans. We compared the performance of our proposed SCSAN model against several established models commonly used in brain tumor classification. These models included pretrained architectures, while the PDCNN model was built

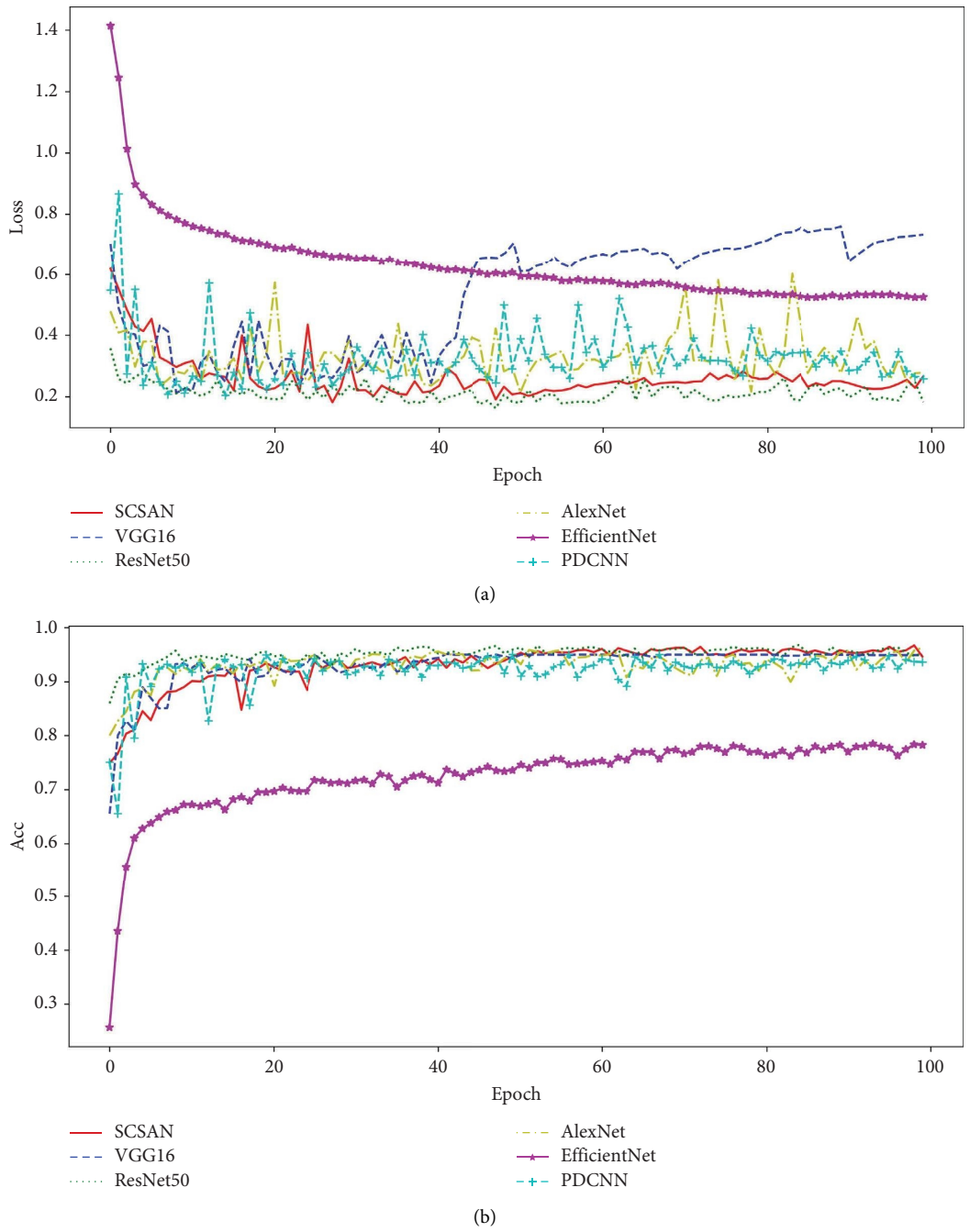

FIGURE 11: Losses and accuracy of the different deep learning models. (a) The x-axis represents an epoch and the y-axis represents a loss. (b) The x-axis represents an epoch and the y-axis represents an accuracy.

from scratch following the author's [17] description. All models were trained for 100 epochs with a learning rate of $1e-4$. Figure 11 illustrates the evolution of test loss and test accuracy for all the models compared to the SCSAN model. In addition, Table 1 presents the overall accuracy, individual class precision, and recall for each model in this experiment. As evident from the results, the SCSAN model achieved the highest accuracy of 0.9644. Notably, the SCSAN model achieves this performance while being significantly lighter in terms of parameters.

4.5. Visualize Saliency Map. In this work, we employed Grad-CAM to generate saliency maps for the proposed SCSAN model as shown in Figure 12 that depicts saliency maps for four input images belonging to distinct classes: no-tumor, glioma, meningioma, and pituitary presented top-to-bottom. A saliency map visually highlights the regions in an input image that contribute most significantly to the predicted class score. Analysis of these saliency maps by domain experts can provide valuable insights into the model's faithfulness and trustworthiness regarding its predictions.

Table 1: The performance of various deep learning models for brain tumor classification.

Model	Parameters (M)	Input	Accuracy	Classwise precision				Classwise recall			
				No-tumor	Glioma	Meningioma	Pituitary	No-tumor	Glioma	Meningioma	Pituitary
VGG16	134.27	$32 \times 3 \times 128 \times 128$	0.88	0.97	0.83	0.76	0.93	0.97	0.81	0.76	0.96
ResNet50	23.52	$32 \times 3 \times 128 \times 128$	0.96	0.99	0.93	0.94	0.98	0.99	0.95	0.93	0.99
AlexNet [29]	57.02	$32 \times 3 \times 128 \times 128$	0.95	0.97	0.93	0.92	0.97	0.98	0.94	0.89	0.99
EfficientNet-B7 [22]	64.05	$32 \times 3 \times 128 \times 128$	0.79	0.91	0.68	0.70	0.86	0.94	0.76	0.59	0.85
PDCNN [17]	3.71	$16 \times 32 \times 32$	0.93	0.98	0.90	0.86	0.99	0.97	0.90	0.88	0.97
SCSAN	11.33	$32 \times 3 \times 128 \times 128$	0.9644	0.99	0.93	0.97	0.97	0.97	0.98	0.91	0.99

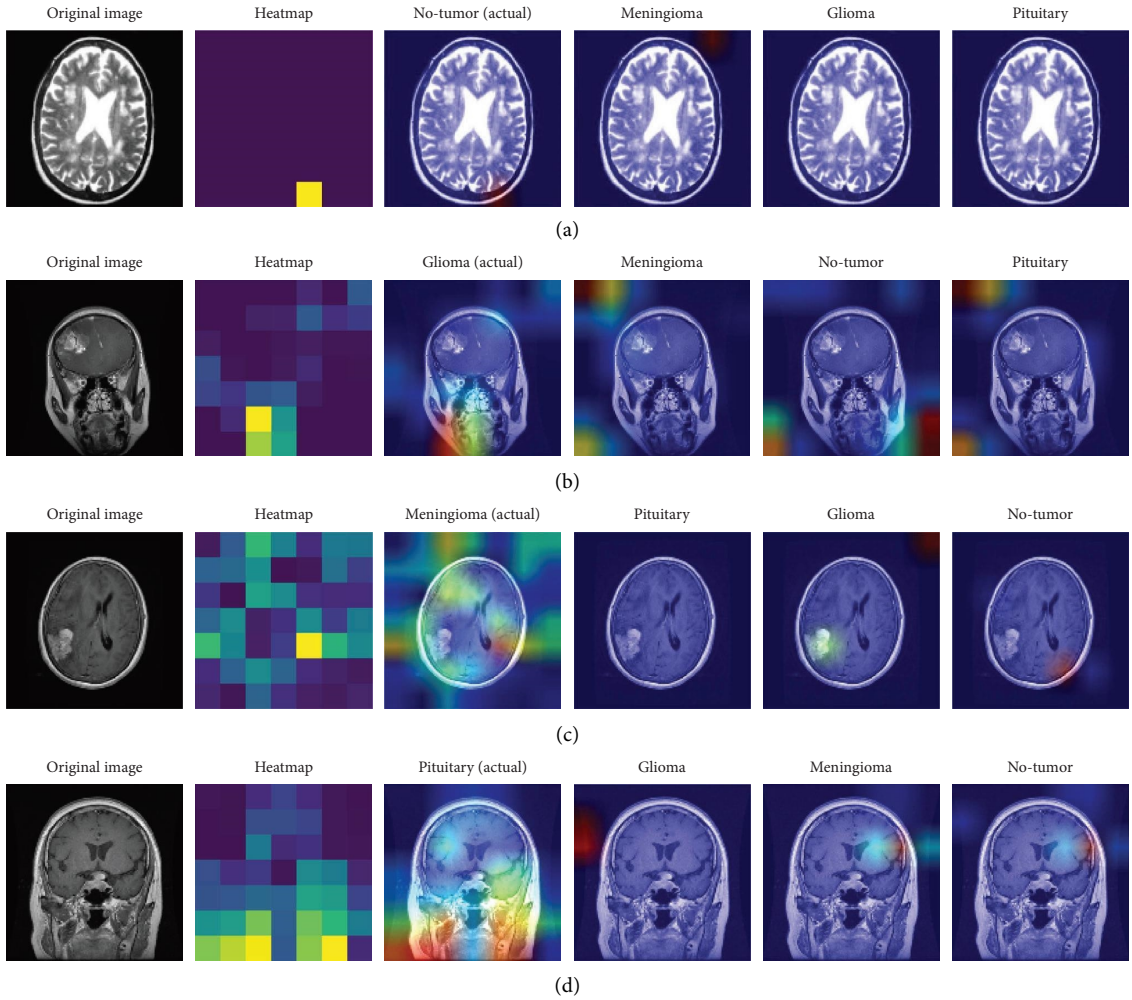

FIGURE 12: The original input image, the corresponding heatmap visualization, and four individual saliency maps. These saliency maps illustrate the regions in the input image that the SCSAN model identifies as most likely to belong to each of the four classes. (a) Analyzing the "no-tumor" class, the SCSAN model focuses on the bottom right pixel of the input image for its prediction. This suggests an absence of prominent features across the entire image that might bias the model towards predicting other classes. (b) In the case of the "glioma" class, the SCSAN model primarily focuses on the bottom middle region of the input image for predicting a glioma tumor. However, the saliency map also highlights areas outside the tumor region, suggesting the model might consider these regions when predicting other classes. (c) The SCSAN model attends to the entire region of the meningioma input image for its prediction as the "Meningioma" tumor class. Notably, the saliency map for other classes lacks prominent features, with the exception of the "glioma" class, where the model focuses on the tumor area. (d) For the "pituitary" class, the SCSAN model primarily focuses on the bottom region of the input image. The saliency map for other classes shows a lack of significant features that might bias the model.

Faithfulness refers to whether the model focuses on the most relevant image regions for classification. Trustworthiness reflects whether these highlighted regions align with human expectations and contribute to accurate predictions. These insights are crucial for determining the model's suitability for real-world applications.

5. Conclusion

The axiom "unity is strength" highlights the effectiveness of combining both local and global features to achieve a high-performance model. In this paper, the convolutional component of the proposed model is responsible for extracting local features and using them for output prediction. Simultaneously, the self-attention component captures global features to enhance the prediction process. The outputs from both components are combined to derive the final result. The proposed model achieved an accuracy of 96.44%, with the highest precision of 99% for the no-tumor class and the lowest precision of 93% for the glioma class. The highest recall was 99% for the pituitary class, while the lowest recall was 91% for the meningioma class. In addition, the visualization showed the regions of the input image that influence the SCSAN model's predictions. The customized Grad-CAM successfully generated heatmaps on the input images, providing insights into the model's decision-making process. Due to the ensembling of two different deep learning models, potential limitations include the risk of

feature redundancy, where extracted features from both models may overlap, leading to inefficiencies. In addition, the model's performance may degrade when processing noisy or low-quality data, as the ensemble approach could amplify irrelevant or misleading features, thereby reducing overall predictive accuracy. To address data privacy concerns, federated learning technology can be employed for deploying the proposed model in brain tumor classification. We also plan to extend the model from 2D to 3D multimodal MRI images to tackle tasks such as segmentation, detection, and classification

Data Availability Statement

The data used to support the findings of this study are included within the article.

Ethics Statement

This research did not contain any studies involving animal or human participants, nor did it take place in any private or protected areas. No specific permissions were required for corresponding locations.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

Rahad Khan is currently doing his Master of Science in Computer Science and Engineering degree under the supervision of Rafiqul Islam who is a faculty member in the department of Computer Science and Engineering at Dhaka University of Engineering & Technology, Gazipur. Rafiqul Islam has analyzed the study and planned the research experiment. Both authors have directly participated in the execution of this work, resulting in and writing the paper equally.

Funding

The authors received no specific funding for this work.

Acknowledgments

The authors thank the Department of Computer Science and Engineering of Dhaka University of Engineering & Technology, Gazipur, for providing research support.

References

[1] M. Arabahmadi, R. Farahbakhsh, and J. Rezazadeh, "Deep Learning for Smart Healthcare: A Survey on Brain Tumor Detection from Medical Imaging," *Sensors* 22, no. 5 (2022): 1960.

[2] A. Verma, S. N. Shivhare, S. P. Singh, N. Kumar, and A. Nayyar, "Comprehensive Review on Mri-Based Brain Tumor Segmentation: A Comparative Study From 2017 Onwards," *Archives of Computational Methods in Engineering* 47 (2024).

[3] Z. Y. Hamd, E. G. Osman, A. I. Alorainy, et al., "The Role of Machine Learning in Detecting Primary Brain Tumors in Saudi Pediatric Patients Through Mri Images," *Journal of Radiation Research and Applied Sciences* 17, no. 3 (2024): 100956.

[4] D. LaBella, U. Baid, O. Khanna, et al., "Analysis of the Brats 2023 Intracranial Meningioma Segmentation Challenge," *arXiv preprint arXiv:2405.09787* (2024).

[5] R. Zhang, H. Luo, W. Chen, Y. Bai, et al., "Review of Deep Learning-Driven Mri Brain Tumor Detection and Segmentation Methods," *Advances in Computer, Signals and Systems* 7, no. 8 (2023): 17–28.

[6] Y. Lee, K. Lee, and S. Pan, "Local and Global Feature Extraction for Face Recognition," in *International Conference on Audio-and Video-Based Biometric Person Authentication* (Springer, 2005), 219–228.

[7] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, and M. Islam, "Combining the Transformer and Convolution for Effective Brain Tumor Classification Using Mri Images," *Applied Sciences* 13, no. 6 (2023): 3680.

[8] S. Reddy, S. Akashdeep, R. Harshvardhan, and S. Kamath, "Stacking Deep Learning and Machine Learning Models for Short-Term Energy Consumption Forecasting," *Advanced Engineering Informatics* 52 (2022): 101542.

[9] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, "An Explainable 3d Residual Self-Attention Deep Neural Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural Mri," *IEEE Journal of Biomedical and Health Informatics* 26, no. 11 (2021): 5289–5297.

[10] M. R. Karim, A. Rahman, and R. Islam, "A Multi-Cancer Detection and Localization System Utilizing X-Ai and Ensemble Technique Using Cnn," in *2024 6th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (2024), 475–480, https://doi.org/10.1109/ICEEICT62016.2024.10534377.

[11] T. T. H. Nguyen, V. B. Truong, V. T. K. Nguyen, Q. H. Cao, and Q. K. Nguyen, "Towards Trust of Explainable Ai in Thyroid Nodule Diagnosis," in *International Workshop on Health Intelligence* (Springer, 2023), 11–26.

[12] L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, "Scouter: Slot Attention-Based Classifier for Explainable Image Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 1046–1055.

[13] Md. Ashafuddula and R. Islam, "Contourtl-Net: Contour-Based Transfer Learning Algorithm for Early-Stage Brain Tumor Detection," *International Journal of Biomedical Imaging* 2024, no. 1 (2024): 6347920.

[14] W. Wenxuan, C. Chen, D. Meng, Y. Hong, Z. Sen, and L. Jiangyun, "Transbts: Multimodal Brain Tumor Segmentation Using Transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2021), 109–119.

[15] A. Anaya-Isaza, L. Mera-Jiménez, L. Verdugo-Alejo, and L. Sarasti, "Optimizing Mri-Based Brain Tumor Classification and Detection Using Ai: A Comparative Analysis of Neural Networks, Transfer Learning, Data Augmentation, and the Cross-Transformer Network," *European Journal of Radiology Open* 10 (2023): 100484.

[16] R. Islam, S. Imran, M. Ashikuzzaman, and M. M. A. Khan, "Detection and Classification of Brain Tumor Based on

Multilevel Segmentation With Convolutional Neural Network," *Journal of Biomedical Science and Engineering* 13, no. 4 (2020): 45–53.

[17] T. Rahman and M. S. Islam, "Mri Brain Tumor Detection and Classification Using Parallel Deep Convolutional Neural Networks," *Measurement: Sensors* 26 (2023): 100694.

[18] D. Maurmo, T. Ruga, E. Zumpano, and E. Vocaturo, "Diagnosing Brain Tumors: An Artificial Intelligence Modeling Approach," *SPAST Reports* 1, no. 4 (2024).

[19] A. Raza, H. Ayub, J. A. Khan, et al., "A Hybrid Deep Learning-Based Approach for Brain Tumor Classification," *Electronics* 11, no. 7 (2022): 1146.

[20] J. Seetha and S. S. Raja, "Brain Tumor Classification Using Convolutional Neural Networks," *Biomedical and Pharmacology Journal* 11, no. 3 (2018): 1457.

[21] A. U. Haq, J. P. Li, S. Khan, M. A. Alshara, R. M. Alotaibi, and C. Mawuli, "Dacbt: Deep Learning Approach for Classification of Brain Tumors Using Mri Data in Iot Healthcare Environment," *Scientific Reports* 12, no. 1 (2022): 15331.

[22] H. M. T. Khushi, T. Masood, A. Jaffar, M. Rashid, and S. Akram, "Improved Multiclass Brain Tumor Detection via Customized Pretrained Efficientnetb7 Model," *IEEE Access* (2023).

[23] M. Aamir, Z. Rahman, Z. A. Dayo, et al., "A Deep Learning Approach for Brain Tumor Classification Using Mri Images," *Computers and Electrical Engineering* 101 (2022): 108105.

[24] R. Chelghoum, A. Ikhlef, A. Hameurlaine, and S. Jacquir, "Transfer Learning Using Convolutional Neural Network Architectures for Brain Tumor Classification from Mri Images," in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2020), 189–200.

[25] C. Srinivas, N. P. Ks, M. Zakariah, et al., "Deep Transfer Learning Approaches in Performance Analysis of Brain Tumor Classification Using Mri Images," *Journal of Healthcare Engineering* 2022 (2022).

[26] Ö.M. Kökçam, A. Boyaci, and M. E. Çolak, "Deep Learning Based Brain Tumor Classification for Mr Images Using Resnet50," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)* (IEEE, 2024), 1–6.

[27] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable Ai for Brain Tumor Detection and Classification," *IEEE Journal of Biomedical and Health Informatics* (2023).

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database" (2009).

[29] M. Siar and M. Teshnehlab, "Brain Tumor Detection Using Deep Neural Network and Machine Learning Algorithm," in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)* (IEEE, 2019), 363–368.

[30] M. O. Khairandish, M. Sharma, V. Jain, J. M. Chatterjee, and N. Jhanjhi, "A Hybrid Cnn-Svm Threshold Segmentation Approach for Tumor Detection and Classification of Mri Brain Images," *Irbm* 43, no. 4 (2022): 290–299.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale" (2020).

[32] F. Shamshad, S. Khan, S. W. Zamir, et al., "Transformers in Medical Imaging: A Survey," *Medical Image Analysis* (2023): 102802.

[33] I. Abedeen, M. A. Rahman, F. Z. Prottyasha, T. Ahmed, T. M. Chowdhury, and S. Shatabda, "Fracatlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs," *Scientific Data* 10, no. 1 (2023): 521.

[34] S. Sangui, T. Iqbal, P. C. Chandra, S. K. Ghosh, and A. Ghosh, "3d Mri Segmentation Using U-Net Architecture for the Detection of Brain Tumor," *Procedia Computer Science* 218 (2023): 542–553.

[35] C. C. Ukwuoma, Z. Qin, M. B. B. Heyat, et al., "A Hybrid Explainable Ensemble Transformer Encoder for Pneumonia Identification From Chest x-Ray Images," *Journal of Advanced Research* 48 (2023): 191–211.

[36] C. Szegedy, W. Liu, Y. Jia, et al., "Going Deeper With Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 1–9.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–778.

[38] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017).

[39] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 10012–10022.

[40] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training Data-Efficient Image Transformers Amp; Distillation Through Attention," *International Conference on Machine Learning* 139 (2021): 10347–10357.

[41] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-Cam: Why Did You Say That? arXiv Preprint arXiv:1611.07450" (2016).

[42] M. Nickparvar, *Brain Tumor MRI Dataset* (Kaggle, 2021), https://www.kaggle.com/dsv/2645886.