Biomedical Signal Processing and Control 112 (2026) 108636

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Enhancing brain tumor classification with a novel attention based explainable deep learning framework

Md Jahid Hasan [a,1], Mahmudul Hasan [b,c,1], Sumya Akter [c], Abu Bakar Siddique Mahi [d], Md Palash Uddin [b,c,*]

[a] Department of Business Information Systems, RMIT University, Melbourne, 3000, Australia
[b] School of Information Technology, Deakin University, Geelong, 3220, Australia
[c] Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, 5200, Bangladesh
[d] Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, 1205, Bangladesh

ARTICLE INFO

Keywords:
Brain tumor classification
Attention network
Strip-Style Pooling Attention
Deep learning
Explainability

ABSTRACT

Accurate and early detection of brain tumors is essential for effective treatment planning in medical diagnosis. However, deep learning (DL) models often struggle with MRI-based tumor detection due to significant variability in tumor size, shape, and location. Traditional diagnostic techniques are limited by subjectivity and low interpretability, while many DL models operate as black boxes, reducing clinical trust. Incorporating attention mechanisms can help by directing the model's focus to the most informative regions of an image, thus improving both accuracy and interpretability. However, existing attention methods often fail to capture the complex spatial and contextual features present in medical images such as MRI scans. In this study, we propose a novel attention-based, explainable DL framework designed to improve the performance and transparency of brain tumor diagnosis. We introduce the Strip-Style Pooling Attention Network (SSPANet), which combines the strengths of channel and spatial attention mechanisms to more effectively capture intricate imaging features. We evaluated SSPANet using VGG16 and ResNet50 as backbone architectures, integrating it alongside existing attention methods for comparison. Among all configurations, ResNet50 combined with SSPANet achieves the best results, with 97% accuracy, precision, recall, and F1-score, along with 95% Cohen's Kappa and Matthews Correlation Coefficient. For interpretability, we employ GradCAM, GradCAM++, and EigenGradCAM across attention-guided DL models. The ResNet50 + SSPANet + GradCAM++ combination consistently provides superior visual explanations, highlighting SSPANet's ability to capture complex spatial-contextual information effectively. We also offer a theoretical analysis to support the efficiency and effectiveness of the proposed attention mechanism.

1. Introduction

A critical disease among the millions of diseases is Brain Tumors (BTs) that develop inside the brain due to the uneven growth of the skull or brain tissue surrounding it [1]. These tumors apply pressure on the brain as they grow erratically throughout the brain. Under pressure, it induces various brain disorders that affect the body, such as dizziness, headaches, fainting attacks, paralysis, etc. This mass of abnormal cells is classified as benign, known as non-cancerous, or malignant, known as cancerous [2].

Studies indicate a growing global incidence of brain tumors. According to estimates, 700,000 Americans are thought to be affected by brain tumors as of 2019, and about 86,000 of these cases received a diagnosis. About 60,800 of these individuals had benign findings, while 26.170 had malignant findings. Among them, 35% of malignant patients survive in the United States (US) [2]. According to the information provided by the editorial board of cancer.net, it is projected that 24,810 people (about 10,530 women and 14,280 men) will receive a brain or spinal cord tumor diagnosis in 2023 in the US. Throughout their lifespans, less than 1% of people will get this type of tumor. In 2020, there will likely be 308,102 new instances of primary brain or spinal cord tumors documented globally [3]. According to the World Health Organization (WHO), 30% of patients with brain tumors do not fully recover from their condition [4]. Additionally, BTs were categorized by the WHO into four grades (Grades I–IV) according to

* Corresponding author at: School of Information Technology, Deakin University, Geelong, 3220, Australia.
E-mail addresses: md.jahid.hasan@rmit.edu.au (M.J. Hasan), mahmudul.hasan@deakin.edu.au (M. Hasan), sumya@hstu.ac.bd (S. Akter), 20101021@uap-bd.edu (A.B.S. Mahi), m.uddin@deakin.edu.au (M.P. Uddin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bspc.2025.108636
Received 3 June 2025; Received in revised form 5 August 2025; Accepted 3 September 2025
Available online 12 September 2025

whether they were benign or malignant. According to the National Brain Tumor Foundation, the number of people who have died from brain tumors over the past three decades has increased by 300% [5].

Brain tumor detection poses a significant challenge in the field of medical diagnosis and treatment, necessitating accurate and efficient detection methods for timely intervention. During the clinical process, medical images are traditionally qualitatively analyzed by skilled medical professionals. However, variations in experience levels among different physicians may introduce biases in the qualitative analysis. The emergence of Deep Learning (DL) algorithms has revolutionized medical image analysis, as they have great learning capabilities, offering promising avenues to improve the precision and efficiency of brain tumor detection [6–10]. The application of DL techniques can aid in clinical processes by mitigating experience discrepancies, minimizing both time and labor expenses, and ultimately improving patient outcomes in neuro-oncology. Its comprehensive overall structure can significantly increase the convenience of practical application for physicians [11,12].

Although DL methodologies have demonstrated remarkable success in various fields, there has been significant interest in the integration of attention mechanisms with DL for medical image analysis in recent years [4,13–15]. These techniques demonstrated the highest level of performance across several DL medical imaging challenges. The attention mechanism draws inspiration from the human cognitive system and functions as a dynamic selection process. Typically, the attention mechanism enhances performance by assigning weights to crucial data components or establishing correlations within the data. An essential factor facilitating the successful integration and utilization of both in medical image analysis is the strong inherent adaptability of the attention mechanism. Furthermore, another advantage of employing attention mechanisms alongside DL for medical images is their potential to enhance the interpretability of DL models. The lesion attention information supplied by the attention mechanism can offer doctors clear and intuitive clinical guidance [11].

Leveraging the power of attention mechanisms, a subset or module of DL holds immense potential for refining the accuracy and interpretability of brain tumor detection systems. The lack of interpretability in the current BT detection system raises significant concerns regarding clinical applicability, safety, and trust. Most existing diagnostic frameworks function as black-box systems, offering high predictive accuracy but failing to provide insights into the reasoning behind their decisions. In high-stakes medical scenarios, such as determining tumor malignancy or delineating treatment plans, this opacity undermines clinician confidence and hinders informed decision-making. For a better understanding of the model's prediction, recently, the integration of Explainable AI (XAI) has become popular in medical diagnosis, which brings trust to the domain experts on AI models' decision-making [13,16,17]. XAI addresses this critical limitation by enabling models to generate human-interpretable justifications for their predictions. Techniques such as SHAP, LIME, and Grad-CAM help identify key imaging features and spatial regions influencing model outcomes, offering transparency that aligns with clinical reasoning. XAI not only facilitates model validation and bias detection but also enhances collaboration between radiologists and AI systems [9]. Moreover, explainability supports regulatory compliance, ethical accountability, and patient trust, essential factors in medical AI deployment.

Although existing DL-based models with attention mechanisms have improved brain tumor diagnosis, existing solutions exhibit several critical limitations. Most current methods incorporate static or heuristic attention modules without adaptively capturing the complex, heterogeneous nature of brain tumors, such as irregular boundaries and varying intensities across modalities. Models like ARM-Net and CKD-TransBTS enhance performance but lack dynamic spatial-contextual integration and remain purely empirically driven [13,18]. Furthermore, the reliance on XAI techniques disconnects interpretability from the core architecture, limiting real-time clinical applicability. Additionally, there is a noticeable absence of theoretical analysis in attention design; no formal understanding of attention behavior or robustness exists across these methods. This highlights the pressing need for a novel, lightweight, and theoretically grounded attention mechanism that provides integrated explainability, spatial precision, and statistical robustness. Such a mechanism should enable both improved classification performance and deeper trust and transparency in high-stakes clinical decision-making environments. Considering all the aspects, this paper presents an exploration into the realm of brain tumor detection using attention-based DL models, aiming to elucidate the benefits, challenges, and prospects of this innovative approach. This paper navigates through recent advancements in attention-based DL techniques, offering insights into their applicability, limitations, and potential avenues for further research. Despite the success of existing attention mechanisms such as Squeeze-and-Excitation (SE), CBAM, and Coordinate Attention in enhancing CNN performance, they suffer from notable limitations.

To address all the limitations, this research work proposes SSPANet, a novel attention mechanism that addresses the limitations of existing methods such as SE, CBAM, and Coordinate Attention. Unlike traditional channel attention techniques that rely solely on global average pooling, SSPANet employs Z-pooling, which combines both average and max pooling to capture richer activation statistics for robust channel-wise modulation. To enhance spatial attention, SSPANet integrates strip pooling, which captures long-range dependencies along orthogonal spatial axes (horizontal and vertical), and style pooling, which incorporates second-order statistics (mean and standard deviation) to model fine-grained textural variations—an aspect overlooked by prior attention modules. These components are fused in a parallel structure with residual learning, enabling efficient and discriminative feature refinement. This unified design allows SSPANet to jointly encode channel saliency, directional spatial context, and style information, making it particularly effective for tasks requiring fine spatial granularity and interpretability, such as medical image classification. The technical contributions of this paper are as follows:

- We design a new attention mechanism, namely the Strip-Style Pooling Attention Network (SSPANet), leveraging the complementary strengths of both channel attention and spatial attention modules for efficient classification of brain tumors from MRI images.
- We introduce a Z-pool-based channel enhancement module in the proposed SSPANet attention mechanism and, for the first time, integrate strip pooling and style pooling to simultaneously capture long-range spatial dependencies and rich texture information. By fusing these strategies, SSPANet significantly enhances spatial awareness while maintaining a low computational footprint.
- We experiment on the FigShare brain tumor multiclass dataset to find the superiority of our proposed attention mechanism. Experimental results confirm that our method shows its superiority and enhances the performance of both baseline VGG16 and ResNet50 without any architectural changes.
- We have employed XAI techniques, GradCAM, GradCAM++, and EigenGradCAM with different combinations of DL models and attention techniques, where the proposed SSPANet shows better transparency, effectiveness, and capability to capture the complex spatial and contextual features in MRI images.
- We provide a theoretical analysis of the proposed SSPANet attention mechanism to justify its effectiveness and efficiency by demonstrating its ability to enhance the signal-to-noise ratio in the channel domain and improve spatial context representation.

The rest of the paper is organized as follows. In Section 2, we present the existing works on tumor detection. In Section 3, we provide the details of the existing methods and our proposed method. The experimental results and analysis are in Section 4. A brief theoretical analysis of the proposed method is presented in Section 5. Finally, in Section 6, we provide the conclusions and future direction of our study.

2. Literature review

Brain tumor detection is one of the more explored domains in computer vision. Already, a lot of work has been completed in this domain, and researchers currently focus on different points of view. In this section, we present some related works relevant to our study.

To detect BTs, [6] proposed an ensemble deep neural support vector machine classifier to classify MRI BT images. They used an adaptive contrast enhancement algorithm and a median filter to preprocess the images, fuzzy c-means to segment the images, and a gray-level co-occurrence matrix to extract the features. In another study, to detect BTs from MRI brain images, including healthy brains, meningiomas, gliomas, and pituitary gland tumors, a convolutional auto-encoder network, a unique 2D CNN architecture, and six widely used ML techniques are used [7]. The authors of the study [19] suggested a modified U-Net structure based on residual networks that employ sub-pixel convolution at the decoder portion and periodic shuffling at the encoder part of the original U-Net. This encoder–decoder structure for brain tumor detection used two MRI datasets, BraTS 2017 and BraTS 2018, to classify tumors effectively. Five well-known DL architectures that make use of the deep Transfer Learning (TL) technique for the multi-class classification of brain tumors [10]. Authos [20] applied four TL algorithms, namely Xception, ResNet50V2, InceptionResNetV2, and DenseNet201, to classify brain tumors, combining preprocessing, reconstruction of TL architectures, and fine-tuning. The proposed method of [21] constructed a DL framework with five-fold cross-validation and a complex multi-branch network with Inception blocks to classify multiclass classification of brain tumors. An efficient technique called SCAOA_DenseNet, which is a combination of the Archimedes Optimization Algorithm and Sine Cosine Algorithm, was developed by [22] for brain tumor detection. They used a Gaussian filter to remove noise, and SegNet, tuned with SCAOA, was used for segmentation. Finally, ShCNN tuned with the proposed SCAOA was used for brain tumor detection, and DenseNet tuned with SCAOA was used for further classification of brain tumors. Another study [23] presented a new automatic segmentation technique called TransU2-Net, which is a combination of a transformer and a lighter U2-Net for multimodal brain tumor segmentation. They claimed that this nested 2D U-shaped network is the first transformer application that has higher accuracy than any 3D medical segmentation application. Authors [24] developed RS2-Net, an efficient multitask learning network that improves tumor classification from medical images by using its predictive mask. This approach is suggested to improve the utilization of the lesion mask and, consequently, the classification performance. To increase the performance of transformer-related brain tumor segmentation with missing modalities, [8] presented a Multiscale Multimodal Vision Transformer (MMMViT). To accommodate brain tumors of different sizes, MMMViT used correlations between modalities and creatively fused local multiscale features to obtain global multiscale features implicitly.

XAI is widely used to interpret the results of the model applied by [16] in the field of brain tumor detection and classification. They used cGAN for image augmentation and four pre-trained models, namely VGGNet, MobileNet, InceptionResNet, and EfficientNet. Again, to ensure the model's transparency, another XAI method, LIME, was employed by [17] for their proposed deep neural network called NeuroNet19, which is a combination of VGG19 and Inverted Pyramid Pooling Module modules for multiscale feature extraction. By integrating a lightweight global attention module with the RM-Net architecture [13] presented ARM-Net, a novel attention-guided residual multiscale CNN designed for multiclass brain tumor classification. The model can selectively learn discriminative features and extract wide-range feature dependencies, leading to improved classification accuracy. The research of [9] generated a Cross-Domain Attention-Guided GAN (CDA-GAN) model, a novel generative data augmentation solution. By generating synthetic samples and balancing class distributions, CDA-GAN effectively tackles limited data challenges in medical imaging. The study demonstrates CDA-GAN's superiority over traditional augmentation methods and GAN-based approaches in classification and segmentation tasks. By integrating clinical knowledge with advanced algorithms, [18] presented a novel method named CKD-TransBTS to diagnose brain tumors from multiple MRI modalities, having two contributions: one is the suggested Modality-Correlated Cross-Attention block, which is incorporated into a dual-branch hybrid encoder that is used to extract the multi-modality image characteristics. Another one is a novel Trans&CNN Feature Calibration, which is suggested as a way to close the disparity and lessen the feature bias between CNN and Transformer. All the studies focus either on performance improvement or explainability, but there is still a lot of scope to improve the model's efficiency and interpretation. Most of the works are based on experimental solutions, which creates a lack of theoretical analysis to create knowledge in this field. It motivates us to provide both the experimental and theoretical results of our proposed method for better clarification and understanding.

3. Methodology

In this section, we provide the details of the dataset, existing DL and attention methods, the descriptions of our proposed SSPANet with detailed architecture, XAI techniques, and performance evaluation methods to get the final decision. We begin with an overview of our proposed methodology with a block diagram, then all the parts are described one by one.

3.1. Approach overview

The proposed solution starts with preprocessing the medical image data, which is then partitioned into training, validation, and testing datasets, which is shown in Fig. 1. We perform BT detection using renowned backbone CNN architectures VGG16 and ResNet50, enhanced with advanced attention mechanisms such as the proposed SSPANet, SRMNet, GCNet, SPNet, and coordinate blocks to improve spatial and contextual representation. To maintain transparency and interpretability of the decisions made by the model, explainable tools, GradCAM, GradCAM++, and EigenGradCAM, are used to focus on the important feature region that has decision-making power for the model. Lastly, the model is evaluated by various metrics such as accuracy, precision, recall, F1-score, Cohen's kappa, MCC, and Matthews Correlation Coefficient (MCC), ensuring a balanced and reliable assessment of performance.

3.2. Descriptions of the dataset

Medical image datasets, such as those involving brain tumors, are often highly imbalanced, with certain classes significantly underrepresented. This imbalance poses challenges for developing reliable AI-based clinical solutions [25] due to biased learning and reduced diagnostic accuracy. In this study, we utilize the Figshare [26] Brain Tumor dataset for experimental analysis, which also exhibits significant class imbalance, necessitating appropriate preprocessing and modeling strategies. The Figshare Brain Tumor dataset comprises 3064 T1-weighted contrast-enhanced MRI images obtained from 233 individuals diagnosed with one of three distinct types of brain tumors: meningioma, glioma, or pituitary tumor. The dataset consists of 708 slices for meningioma, 1426 slices for glioma, and 930 slices for pituitary tumors. Each image has a resolution of 512 pixels.

A total of 2451 images (80%) have been allocated for training, providing the model with substantial exposure to a diverse range of tumour types. An additional 306 images (10%) form the validation set, used to fine-tune model parameters and monitor performance on unseen data, thereby mitigating the risk of overfitting. The remaining 307 images (10%) constitute the test set, reserved exclusively for evaluating the model's generalization capability on previously unobserved cases. Table 1 presents the distribution of images across the three subsets.

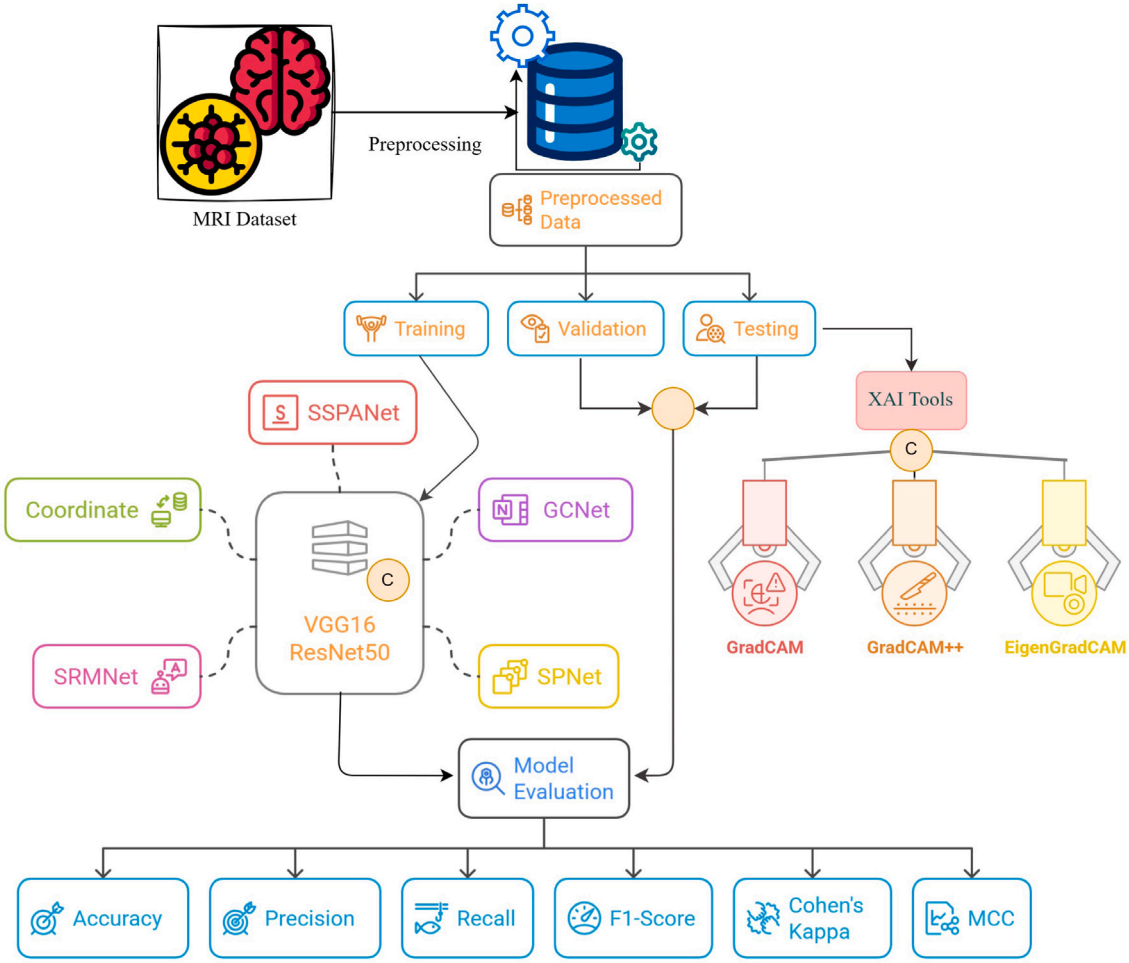

Fig. 1. Overview of the proposed brain tumor detection method. It includes basic preprocessing, model development with attention integrations, and explainability on all combinations (C).

Table 1
Proportional distribution of images by class and subset in the Brain Tumor dataset.

Tumour type	Train (80%)	Val (10%)	Test (10%)
Meningioma	566	71	71
Glioma	1140	143	143
Pituitary	744	93	93
Total	2450	307	307

3.3. Descriptions of benchmark DL models

We use the VGG16 and ResNet50 benchmark DL models as the backbone of this analysis. The working process and architectural descriptions of the models are below.

3.3.1. VGG16

VGG16 consists of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers [27,28]. Every convolutional layer applies 3×3 filters with a stride value of 1, followed by a ReLU activation. The max-pooling layers achieve down-sampling by using filters of 2×2 size and a stride of 2 to decrease feature map dimensions. The output of convolutional layers follows a flattening process that leads to fully connected layers before a softmax function performs the classification. The convolution operation is defined as Eq. (1).

$$Y_{i,j,k} = \sum_{m=0}^{2} \sum_{n=0}^{2} \sum_{c=0}^{C-1} X_{i+m,j+n,c} \cdot W_{m,n,c,k} + b_k \tag{1}$$

This equation describes how each output element $Y_{i,j,k}$ is computed by applying a 3×3 filter W to the input X and adding a bias term b_k. The ReLU activation function is expressed as Eq. (2).

$$f(x) = \max(0, x) \tag{2}$$

Here, all negative values in the feature map are set to zero, introducing non-linearity. The max-pooling is represented as Eq. (3).

$$Y_{i,j} = \max\left(X_{i,j}, X_{i+1,j}, X_{i,j+1}, X_{i+1,j+1}\right) \tag{3}$$

This operation selects the maximum value from a 2×2 window, down-sampling the feature map. Fully connected layers that are computed as Eq. (4).

$$y = W \cdot x + b \tag{4}$$

where x is the input vector, W is the weight matrix, and b is the bias term, producing the output y. The softmax output function for classification is defined as Eq. (5).

$$P(y = c \mid x) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \tag{5}$$

This equation converts the output logits z from the final fully connected layer into a probability distribution over the C classes.

3.3.2. ResNet50

ResNet50, consisting of 50 layers, implements residual blocks to enable the training of very deep networks [29,30]. Here, each residual block includes convolutional layers, batch normalization, and ReLU

activation functions. The operation of a residual block can be expressed as Eq. (6)

$$\mathcal{F}(x) = \text{ReLU}\left(\text{BN}\left(W_2 \cdot \text{ReLU}\left(\text{BN}\left(W_1 \cdot x + b_1\right)\right) + b_2\right)\right) \tag{6}$$

where BN refers to Batch Normalization, W_1, W_2 are the weights of the first and second convolutional layers, b_1, b_2 are the biases for the convolutional layers, x is the input to the residual block, and $\mathcal{F}(x)$ is the output of the residual function. The output of each residual block is obtained by adding the input x to the residual function output $\mathcal{F}(x)$ via the skip connection, as in Eq. (7).

$$y = \mathcal{F}(x) + x, \tag{7}$$

where y is the final output of the residual block. This skip connection helps to avoid the vanishing gradient problem and ensures better gradient flow through the network, enabling the training of deeper architectures.

3.4. Existing attention networks

3.4.1. Coordinate

The coordinate attention mechanism consists of two key steps: coordinate information embedding and coordinate attention generation [31]. First, instead of applying global pooling across both spatial dimensions, two spatial extents of pooling kernels encode each channel horizontally and vertically as Eq. (8).

$$f_c^h(i) = \frac{1}{W}\sum_{j=1}^{W} X_c(i,j), \quad f_c^w(j) = \frac{1}{H}\sum_{i=1}^{H} X_c(i,j) \tag{8}$$

where $f_c^h(i)$ and $f_c^w(j)$ represent the horizontally and vertically aggregated features, respectively. These pooled features are concatenated and transformed using a shared 1×1 convolution as in Eq. (9).

$$z = \delta(W_1 \cdot [f^h, f^w]) \tag{9}$$

where W_1 is a learnable transformation matrix and δ represents a nonlinear activation function. The resulting tensor is then split into two attention vectors for horizontal and vertical coordinate encoding as Eq. (10).

$$a^h = \sigma(W_h \cdot z), \quad a^w = \sigma(W_w \cdot z) \tag{10}$$

where W_h and W_w generate position-sensitive attention maps. Finally, these attention vectors modulate the input feature map as in Eq. (11).

$$\tilde{X}_c(i,j) = X_c(i,j) \cdot a^h(i) \cdot a^w(j) \tag{11}$$

This equation represents the final recalibration step in the Coordinate Attention mechanism. Here, the original feature $X_c(i,j)$ at channel c, row i, and column j is scaled by two attention weights: $a^h(i)$, which captures long-range dependencies along the height, and $a^w(j)$, which encodes spatial correlations along the width. By applying these attention maps independently in two spatial directions, Coordinate Attention allows the network to emphasize important regions while maintaining positional information and minimizing computational overhead.

3.4.2. GCNet

GCNet is a cutting-edge global context network that revolutionizes global context modeling in vision tasks such as object detection, instance segmentation, image classification, and action recognition [32]. It introduced Global Context (GC) blocks, which are lightweight and efficient at capturing long-range dependencies in images. GC blocks are applied to multiple layers to create GCNet. The architecture of the GC block is formulated as Eq. (12).

$$z_i = x_i + W_{v2} ReLU(LN(W_{v1} \sum_{j=1}^{\infty} \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j)), \tag{12}$$

where $\alpha_j = \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j$ denotes the weight of global attention pooling. $\delta(.) = x_i + W_{v2} ReLU(LN(W_{v1} \sum_{j=1}^{\infty}(.)))$ denotes the bottleneck transform.

The GC block comprises global attention pooling for context modeling, a bottleneck transform to capture dependencies across channels, and broadcast element-wise addition for feature fusion. With only a small increase in calculation cost, the GC block can be used in numerous layers to capture the long-range dependency better. This characteristic makes this attention lightweight.

3.4.3. SPNet

Spatial pooling typically operates within a small region, limiting its ability to capture long-range dependencies and focus on distant regions. To address this, Hou et al. proposed strip pooling, a novel pooling method that is a lengthy but narrow pooling window that enables the model to gather rich global contextual data [33]. They introduced a novel Strip Pooling Module (SPM) and a Mixed Pooling Module (MPM) to produce superior segmentation forecasts. The SPM gathers long-range context across multiple spatial dimensions by utilizing both horizontal and vertical strip pooling processes. The input feature map is denoted as $X \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, H is the height (number of rows), and W is the width (number of columns). Horizontal strip pooling operates by pooling across the width (columns) of the feature map for each row. The output of horizontal strip pooling for a given feature map X is computed as Eq. (13).

$$S_h(X)_{c,i} = \frac{1}{W}\sum_{j=1}^{W} X_{c,i,j} \tag{13}$$

where $S_h(X)_{c,i}$ is the pooled value for the cth channel and ith row. Vertical strip pooling is computed by pooling across the height (rows) of the feature map for each column. The output of vertical strip pooling is computed as Eq. (14).

$$S_v(X)_{c,j} = \frac{1}{H}\sum_{i=1}^{H} X_{c,i,j} \tag{14}$$

where $S_v(X)_{c,j}$ is the pooled value for the cth channel and jth column. The SPM gathers long-range context from both the horizontal and vertical directions independently, leading to richer feature representations. The MPM aims to increase the discriminativeness of feature representations by aggregating diverse contextual information using a variety of pooling techniques.

3.4.4. SRMNet

The Style-based Recalibration Module (SRM) is a straightforward but powerful architectural unit that uses the styles of intermediate feature maps to adaptively recalibrate them [34]. Style integration and style pooling are the two primary parts of SRM. Using style pooling, SRM first collects the style information from every single channel of the feature maps. Then, style integration is used to estimate the recalibration weight for each channel utilizing style features. In the end, the style weights adjust the feature maps to make their information more or less prominent. Thus, it helps to increase the representational power of the CNN. SRM creates channel-wise recalibration weights $G \in \mathbb{R}^{N \times C}$ based on the styles of an input tensor $X \in \mathbb{R}^{N \times C \times H \times W}$. Here, N represents the number of samples in a mini-batch, C is the number of channels, and $H \times W$ denotes the spatial dimensions of the feature maps. To extract style features, SRM computes the channel-wise mean and standard deviation of feature maps, which capture the global statistics of each channel as Eqs. (15) and (16).

$$\mu_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{c,i,j} \tag{15}$$

$$\sigma_c = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(X_{c,i,j} - \mu_c)^2} \tag{16}$$

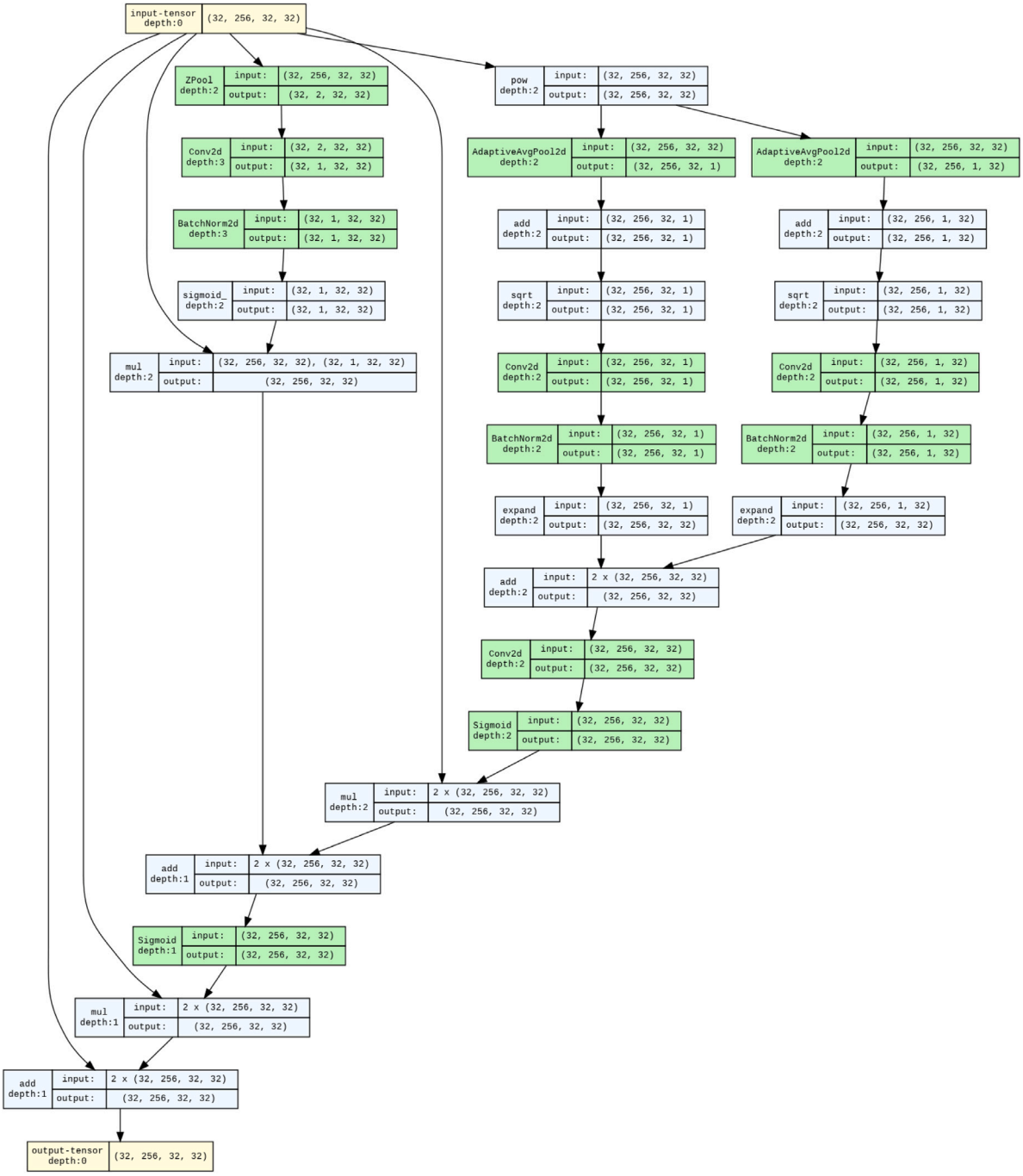

Fig. 2. Proposed SSPANet's architectural mapping between "Z-pooling-driven channel attention" and "directionally aware spatial attention" fused with "style statistics".

Here, μ_c represents the mean of the activations for channel c, summarizing the average intensity of features across spatial locations. σ_c is the standard deviation, which measures the spread of feature values and captures texture-related information. These statistical representations form the style descriptor of each channel and are later used for recalibrating the feature maps.

3.5. Proposed SSPANet

We propose a hybrid attention mechanism, namely SSPANet, to enhance deep convolutional representations by selectively emphasizing informative features in both channel and spatial dimensions. Compared to traditional attention modules that rely solely on global pooling or simplistic context fusion, SSPANet introduces a novel architectural mapping between "Z-pooling-driven channel attention" and "directionally aware spatial attention" fused with "style statistics", as illustrated in Figs. 2 and 3.

Fig. 2 presents the detailed architecture of the proposed hybrid attention mechanism employed in SSPANet. The module receives an intermediate feature map from the CNN backbone and processes it through parallel spatial and channel attention branches. The channel branch, located on the left, applies Z-pooling to compress spatial information by aggregating both average and max features across the channel dimension. A 1×1 convolutional layer is then applied, followed by batch normalization and a sigmoid activation, producing an attention map that is broadcast and multiplied by the input tensor.

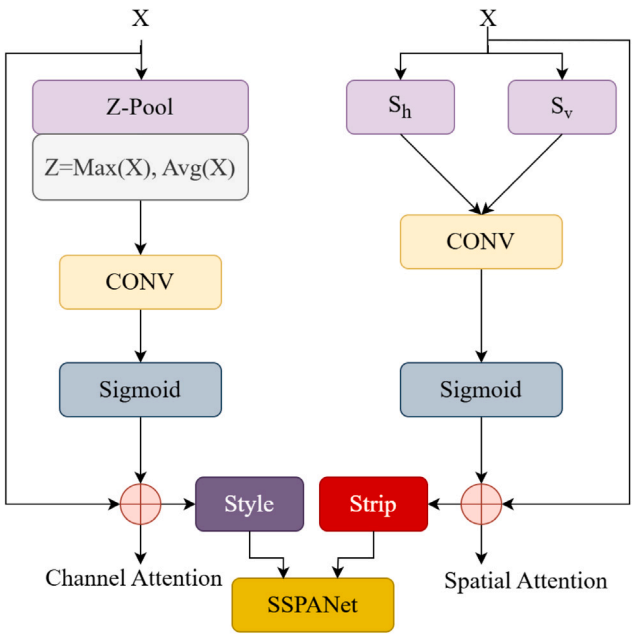

Fig. 3. Structural diagram of proposed SSPANet.

This step adjusts weights across feature channels based on importance, thereby enhancing spatially informative regions.

The middle and right branches implement directionally aware spatial attention. Both branches first perform adaptive average pooling along the vertical and horizontal spatial axes to capture positional relationships in the feature map. To improve numerical stability, the square root of the pooled features is taken after the addition of a small constant. These features are passed through 1×1 convolutional layers and batch normalization, then reshaped (expanded) back to the input tensor's dimensions and added together to form an intermediate representation that captures long-range dependencies across both spatial directions.

Finally, SSPANet fuses the outputs of all three branches along with the original input feature map using a residual connection. This fusion allows the network to jointly exploit channel-wise attention, directional spatial context, and style-aware statistics, resulting in enhanced feature representation and improved classification performance.

To present the working process of SSPANet, consider an input tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the SSPANet module applies two sequential attention operations *Channel Attention Module* followed by *Spatial Attention Module* to generate a refined feature map $\mathbf{Y}$ as in Eq. (17).

$$\mathbf{Y} = \mathbf{M}_s \otimes (\mathbf{M}_c \otimes \mathbf{X}) \tag{17}$$

Here, $\mathbf{M}_c \in [0,1]^{C\times 1\times 1}$ denotes the channel attention map and $\mathbf{M}_s \in [0,1]^{1\times H\times W}$ represents the spatial attention map, both learned adaptively from the input.

To build the inter-channel dependencies, we first construct a compact channel descriptor using "Z-Pooling", which concatenates the average and max pooling across spatial dimensions as Eq. (18).

$$\mathbf{Z} = \left[\text{AvgPool}_{H,W}(\mathbf{X}); \text{MaxPool}_{H,W}(\mathbf{X})\right] \in \mathbb{R}^{2\times C\times 1\times 1} \tag{18}$$

This pooled representation is processed through a convolutional layer followed by batch normalization and two successive sigmoid activations as in Eq. (19).

$$\mathbf{F}_1 = \text{BN}(\text{Conv2D}(\mathbf{Z})), \quad \mathbf{M}_c = \sigma(\mathbf{F}_1) \tag{19}$$

The resulting $\mathbf{M}_c$ is broadcast and multiplied with the original feature map $\mathbf{X}$, yielding the intermediate output $\mathbf{X}' = \mathbf{M}_c \otimes \mathbf{X}$, where $\otimes$ denotes element-wise multiplication.

The Spatial Attention Module in SSPANet incorporates *strip pooling* for directional awareness and *style pooling* for texture discrimination, in contrast to traditional spatial attention mechanisms that use global pooling or convolutional attention. We use strip pooling along height and breadth, respectively, to extract long-range horizontal and vertical context as Eq. (20).

$$\begin{aligned} \mathbf{S}_h &= \text{AvgPool}_H(\mathbf{X}') \in \mathbb{R}^{C\times 1\times W} \\ \mathbf{S}_v &= \text{AvgPool}_W(\mathbf{X}') \in \mathbb{R}^{C\times H\times 1} \end{aligned} \tag{20}$$

These are concatenated and passed through a convolutional layer as Eq. (21).

$$\mathbf{S}_{\text{strip}} = \text{Conv}_{1\times 1}([\mathbf{S}_h; \mathbf{S}_v]) \tag{21}$$

To design textural consistency, we compute the *mean* and *standard deviation* of channel responses at each spatial location as Eq. (22).

$$\mu(x_{hw}) = \frac{1}{C}\sum_{c=1}^{C} x_{c,hw}, \quad \sigma(x_{hw}) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(x_{c,hw} - \mu)^2} \tag{22}$$

A parallel channel is employed to project and fuse these descriptors. The final spatial attention map is created by adding the outputs from the strip and style branches and running them through a sigmoid activation as in Eq. (23).

$$\mathbf{M}_s = \sigma\left(\text{Conv}_{1\times 1}(\mathbf{S}_{\text{strip}} + \mathbf{S}_{\text{style}})\right) \tag{23}$$

Finally, $\mathbf{M}_s$ is broadcast over the channel dimension and multiplied with $\mathbf{X}'$ to produce the final attention-refined output $\mathbf{Y}$.

As visualized in the computational graph in Fig. 2, the entire SSPANet mechanism is implemented as a directed acyclic flow of operations involving pooling, element-wise arithmetic, 1D and 2D convolutions, and nonlinear activations. Notably, each operation maintains the dimensional consistency of the input $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ across the attention pipeline, ensuring seamless integration into existing convolutional backbones.

The proposed SSPANet shows various architectural improvements over previous attention modules like SE, CBAM, and Coordinate Attention. Unlike typical pooling techniques, strip pooling directly preserves global context along orthogonal spatial axes. Standard deviation, a second-order statistic, is introduced by style pooling to capture subtle changes in look. Strip pooling and lightweight 1×1 convolutions guarantee computational tractability even with increased modeling capacity. The model's performance is experimentally improved when channel attention is applied before spatial attention. This allows the model to first recalibrate feature importance before concentrating on spatial locality. Particularly in tasks where *fine spatial granularity*, *texture consistency*, and *global context* are crucial, these improvements allow SSPANet to learn more resilient, discriminative, and context-aware feature representations.

3.6. Experimental settings

The experimental setup has been carefully configured to ensure reliable and reproducible model training and evaluation. The model was trained for 30 epochs, which was empirically determined to provide sufficient convergence while avoiding overfitting. The training objective was optimized using the cross-entropy loss function, suitable for multi-class classification tasks, and the Adam optimizer was employed with a fixed learning rate of 0.0001, offering adaptive gradient updates that enhance training stability. A batch size of 32 was selected to balance computational efficiency and gradient consistency. Early stopping was applied based on validation loss to prevent overfitting and to retain the model with the best generalization performance.

Model evaluation was conducted using a comprehensive suite of metrics to capture various aspects of classification performance. These included standard statistical metrics such as accuracy, precision, recall, and F1-score, as well as two agreement-based metrics used to assess classification reliability: Cohen's Kappa, which accounts for agreement occurring by chance, and the Matthews Correlation Coefficient (MCC), which provides a balanced measure of performance in the presence of class imbalance.

3.7. Performance measure metrics

Accuracy: Accuracy is the proportion of correctly classified instances among the total instances.

$$\text{Accuracy:} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i + FN_i + TN_i)} \tag{24}$$

where TP_i is the number of true positives for class i, FP_i is the number of false positives for class i, FN_i is the number of false negatives for class i, and TN_i is the number of true negatives for class i.

Precision: Precision is the proportion of true positive instances among the instances classified as positive.

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{25}$$

For multiclass, the average precision can be calculated as the macro-average (average of the precisions of all classes) or the weighted average (weighted by the number of instances per class).

Recall: Recall is the proportion of true positive instances among the actual positives.

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{26}$$

For multiclass, the average recall can be calculated as the macro-average or the weighted average.

F1 Score: F1 score is the harmonic mean of precision and recall, providing a balance between the two.

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{27}$$

For multiclass, the average F1 score can be calculated as the macro-average or the weighted average.

Cohen's Kappa: Cohen's Kappa measures the agreement between two raters (or classifications), taking into account the possibility of agreement occurring by chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{28}$$

where p_o is the observed agreement (accuracy), and p_e is the expected agreement by chance.

$$p_e = \sum_{i=1}^{n} \left(\frac{(TP_i + FP_i)(TP_i + FN_i)}{N^2} \right), \tag{29}$$

where N is the total number of instances.

Matthews Correlation Coefficient (MCC): MCC is a measure of the quality of binary classifications, considering all four confusion matrix categories (TP, TN, FP, FN). It can be extended to multiclass classification.

$$MCC = \frac{c \cdot s - \sum_k p_k \cdot t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \tag{30}$$

where c is the number of correct predictions, s is the total number of samples, p_k is the number of times class k is predicted, and t_k is the number of times class k is true.

3.8. Explainable AI techniques

To provide a comprehensive and complementary perspective on the model's decision-making process, we employed three GradCAM-based explainability techniques: GradCAM, GradCAM++, and Eigen-GradCAM. While tools like SHAP and LIME are powerful for feature-level interpretability [35], they have notable limitations when applied to medical image analysis. Specifically, their perturbation-based approaches lack spatial coherence and do not provide localized visual explanations aligned with clinically meaningful regions. In contrast, GradCAM generates class-discriminative localization maps by utilizing gradient information from the final convolutional layers, offering insight into the regions most influential to the model's prediction. GradCAM++ improves upon this by enabling more precise localization, particularly in scenarios involving multiple or overlapping tumor regions, common in complex medical imaging tasks. In contrast, Eigen-GradCAM applies eigenvalue decomposition on the activation maps to extract principal components, producing noise-reduced, eigenvector-based heatmaps that emphasize dominant visual patterns while suppressing less relevant features. This combination of complementary techniques forms a robust interpretability framework that enhances transparency and diagnostic confidence by cross-validating the focus regions across distinct attention mechanisms. The details of the XAI techniques are below.

GradCAM: Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely utilized method in explainable XAI that facilitates the visual understanding of predictions generated by DL models [36]. This approach identifies and highlights certain regions within an input image that significantly influence the model's final classification result. The approach involves calculating the gradients of the output score for the chosen class about the feature maps of the final convolutional layer, as specified in Eq. (31).

$$\alpha_k^c = \frac{1}{N} \sum_m \sum_n \left(\frac{\partial S^c}{\partial F_{mn}^k} \right), \tag{31}$$

where α_k^c denotes weight for the kth feature map corresponding to class c, S^c represents score for class c, F_{mn}^k denotes activation at spatial location (m, n) in the kth feature map, and N denotes normalization factor (total number of spatial positions).

GradCAM++: Unlike Grad-CAM, which only utilizes the positive gradients of the target class score concerning the feature maps, Grad-CAM++ integrates both positive and negative gradients obtained from the final convolutional layer [37]. This broader gradient usage results in more precise and detailed heat maps, thus improving the interpretability of CNN predictions. Mathematically, Grad-CAM++ computes a weighted sum of higher-order partial derivatives, enabling a more refined visualization of image regions contributing to the model's output decision. The weighting factor is defined by Eq. (32).

$$\alpha_k^y = \sum_i \sum_j w_{ij}^{ky} \cdot \text{ReLU}\left(\frac{\partial Y^y}{\partial A_{ij}^k} \right), \tag{32}$$

where α_k^y represents the weight assigned to the kth feature map for class y, w_{ij}^{ky} indicates the importance of spatial coordinates (i, j) in the kth feature map concerning class y, Y^y denotes the output score corresponding to class y, A_{ij}^k refers to the activation at position (i, j) in the kth feature map.

EigenGrad-CAM: EigenGrad-CAM is a gradient-based visual explanation technique that improves the spatial localization and class-discriminative properties of CNN interpretability methods [38]. Unlike Grad-CAM, which uses globally averaged gradients to weight feature maps, EigenGrad-CAM performs Principal Component Analysis (PCA) on gradient-weighted feature maps to identify the most informative spatial patterns.

Let $A \in \mathbb{R}^{K \times H \times W}$ denote the set of feature maps from a convolutional layer, and $\frac{\partial y^c}{\partial A} \in \mathbb{R}^{K \times H \times W}$ be the gradients of the output score for

Table 2
Performance of different attention mechanisms with VGG16.

Model	Accuracy	Precision	Recall	F1 score	Cohen Kappa	MCC
VGG16	0.89	0.89	0.89	0.89	0.82	0.82
VGG Coordinate	0.90	0.90	0.90	0.90	0.83	0.83
VGG GCNet	0.55	0.55	0.55	0.55	0.00	0.00
VGG SPNet	0.89	0.89	0.89	0.89	0.81	0.82
VGG SRMNet	0.90	0.90	0.90	0.90	0.83	0.83
VGG SSPANet	0.93	0.93	0.93	0.93	0.88	0.89

Table 3
Performance of different attention mechanisms with ResNet50.

Model	Accuracy	Precision	Recall	F1 score	Cohen Kappa	MCC
ResNet	0.90	0.90	0.90	0.90	0.83	0.83
ResNet Coordinate	0.88	0.88	0.88	0.88	0.81	0.81
ResNet GCNet	0.84	0.84	0.84	0.84	0.74	0.74
ResNet SPNet	0.93	0.93	0.93	0.93	0.89	0.89
ResNet SRMNet	0.93	0.93	0.93	0.93	0.88	0.88
ResNet SSPANet	0.97	0.97	0.97	0.97	0.95	0.95

class c concerning these maps. The element-wise product is computed as Eq. (33).

$$G^k = A^k \cdot \frac{\partial y^c}{\partial A^k}, \quad \forall k \in \{1, \ldots, K\} \tag{33}$$

Each $G^k \in \mathbb{R}^{H \times W}$ is flattened and stacked to form the matrix $G \in \mathbb{R}^{K \times (H \cdot W)}$. PCA is then applied to G, and the first principal component $v_1 \in \mathbb{R}^{H \cdot W}$ is reshaped to match the original spatial dimensions as in Eq. (34).

$$L^c_{\text{EigenGrad-CAM}} = \text{reshape}(v_1) \in \mathbb{R}^{H \times W} \tag{34}$$

4. Result analysis

This study evaluates two DL models, VGG16 and ResNet50, each enhanced with various attention mechanisms. The following sections present the experimental results, supported by relevant tables and figures, along with a comparative analysis. Section 4.1 covers attention network performance with VGG16, and Section 4.2 with ResNet50. Section 4.3 explores model explainability in brain tumor detection, while Section 4.4 highlights the superior performance of the proposed SSPANet over other attention mechanisms.

4.1. Performance of the attention networks with VGG16 model

In the first evaluation using the VGG architecture, different attention mechanisms are applied to observe their impact on model performance. We tabulate the results in Table 2. The baseline VGG16 model achieves an accuracy, precision, recall, and F1 score of 89%, with CK and values of 82%. Among the attention-based variants, VGG Coordinate and VGG SRMNet slightly improve performance to 90% across the primary metrics, with CK and MCC rising to 83%. VGG SPNet maintains similar results to the baseline, while VGG GCNet significantly underperforms, scoring only 55% in all key metrics, suggesting poor integration with the VGG backbone. In contrast, the proposed VGG SSPANet achieves the best performance, with a notable increase to 93% in all core metrics and improvements in CK 88% and MCC 89%. This clearly indicates that the SSPANet attention mechanism enhances the VGG model's ability to extract meaningful features and improves overall classification performance. The performance of the models is visualized in Fig. 4, where we present the accuracy and loss graphs of different models in the training and validation phases. The graphs also indicate that integration of the proposed SSPANet with the VGG backbone increases the model performance and outperforms other attention-integrated VGG models and the baseline VGG16.

4.2. Performance of the attention networks with ResNet50 model

In the same experimental setup, we evaluate the ResNet50 architecture integrating different attention mechanisms, and the results are in Table 3. The baseline model shows a good performance with 90% across all major metrics, CK, and MCC values of 83%. ResNet Coordinate and GCNet show comparatively low performance, with GCNet showing only 84% in accuracy, precision, recall, and F1 score, along with 74% in CK and MCC. In contrast, ResNet, SPNet, and SRMNet each achieve 93% in all performance measures, indicating notable enhancement over the baseline. However, the most impressive results were observed with the proposed ResNet SSPANet model, which reached 97% across accuracy, precision, recall, and F1 score, with CK and MCC both climbing to 95%. This highlights the significant contribution of SSPANet in improving model performance. We also show the training, validation accuracy, and loss curve of all the models in Fig. 5. The curves clearly indicate that the performance of the ResNet SSPANet combination is significantly better than other models. It shows a stable training process with high detection performance, making it superior among all the combinations.

The proposed SPANet shows better performance in both architectures. It is evident that the proposed SSPANet attention mechanism consistently delivers superior results across all combinations. It shows performance improvement in both VGG and ResNet models; the gains are more substantial with the deeper ResNet50 architecture, indicating better synergy with complex feature extraction. In both scenarios, SSPANet outperformed all other attention mechanisms and baseline models, demonstrating its robustness, effectiveness, and potential as a powerful attention module in DL pipelines. We show its superiority by a theoretical analysis in the following section.

4.3. Statistical significant test

Table 4 presents the results of paired t-tests comparing the performance of SSPANet with other attention mechanisms using VGG16 and ResNet50 architectures. Each comparison is based on five independent runs. The table reports the t-statistic, p-value, and whether the difference is statistically significant ($p < 0.05$). The results confirm that SSPANet significantly outperforms all baseline methods in both architectures, demonstrating its consistent effectiveness across different backbone networks.

4.4. Explainability results of models in BT detection

In this section, we provide the detailed results of the XAI on BTs detection. To evaluate the effectiveness of the proposed SSPANet, we utilize VGG16 and ResNet50 and apply XAI tools on both models individually. Our main target is to find the efficacy of SSPANet on glioma, meningioma, and pituitary types of tumors, focusing on the features that are most important for classification. We experiment with GradCAM, GradCAM++, and EigenGradCAM with VGG16 and ResNet50 separately and compare their results to find the most accurate pipeline. We apply these three XAI tools with all attention techniques (Coordinate Attention, SRMNet, GCNet, and SPNet) and compare their results with the proposed SSPANet's explainability. Additionally, we compare the baseline VGG16 and ResNet50 explainability with SSPANet.

We generate heatmaps from the VGG16 + SSPANet and ResNet50 + SSPANet combinations shown in Figs. 6 and 7, respectively. These visualizations provide better tumor region identification because they demonstrate decreased noise levels and better spatial accuracy. All three explainability techniques show accurate and delineated activation regions when ResNet50 uses SSPANet for glioma classification. GradCAM++ produces focused attention around tumors that separates lesions from their surrounding environment. In the same way, SSPANet

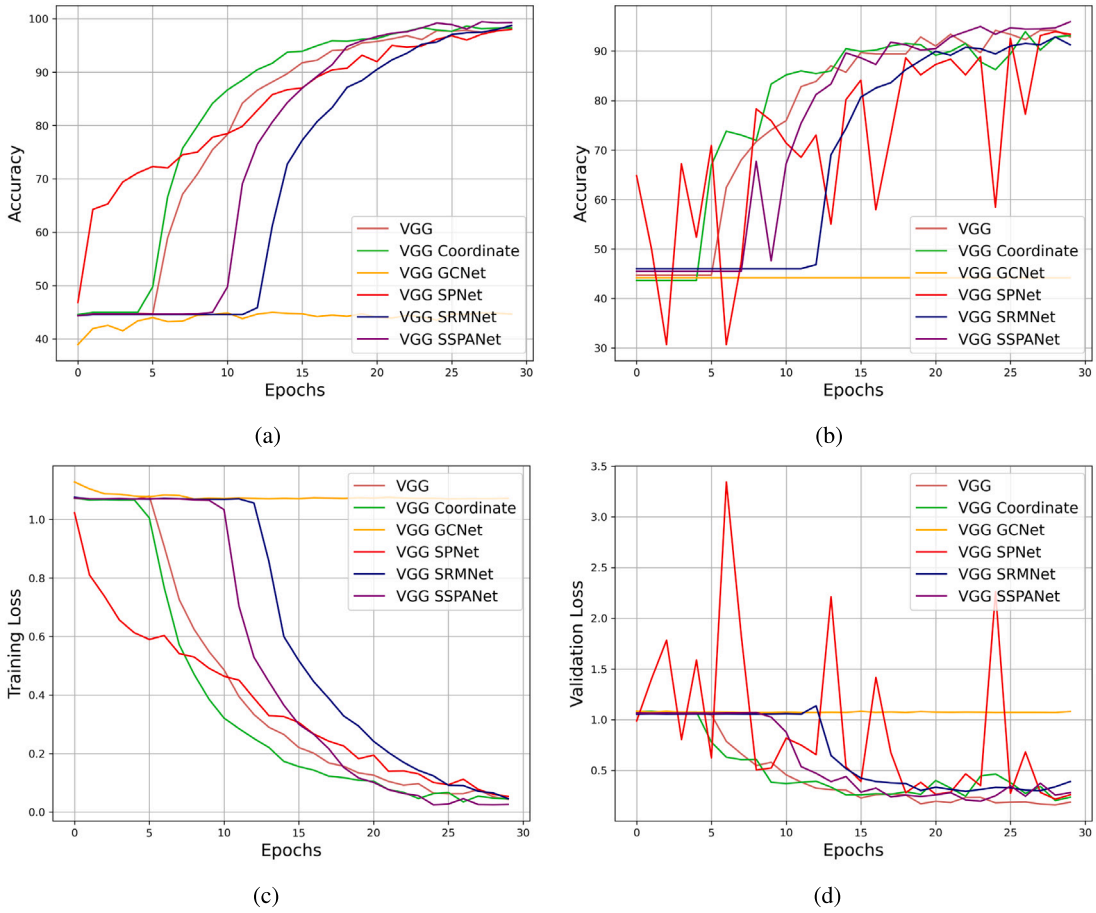

Fig. 4. Accuracy and loss of different attention networks with VGG16. (a) Training Accuracy and, (b) Validation Accuracy (c) Training Loss (d) Validation Loss.

Table 4
Paired t-test results comparing SSPANet with other attention mechanisms on VGG16 and ResNet50 architectures.

ResNet				VGG			
Model	t-statistic	p-value	Significance	Model	t-statistic	p-value	Significance
SRMNet	8.944	0.0009	$p < 0.05$	SRMNet	9	0.0008	$p < 0.05$
SPNet	16	0.0001	$p < 0.05$	SPNet	7.667	0.0016	$p < 0.05$
Coordinate	28.46	0	$p < 0.05$	Coordinate	4.824	0.0085	$p < 0.05$
GCNet	41.11	0	$p < 0.05$	GCNet	120.167	0	$p < 0.05$
ResNet	15.652	0.0001	$p < 0.05$	ResNet	7.667	0.0016	$p < 0.05$

with VGG16 generates exceptional activation clarity through Eigen-GradCAM visualization because it creates maps that perfectly identify tumors with low rates of incorrect detections. The combination of SSPANet with a backbone leads to superior results compared to the base models in meningioma and pituitary classification. The combination of VGG16 + SSPANet with GradCAM++ activates precisely around the tumor area during meningioma detection, while ResNet50 + SSPANet with EigenGradCAM generates activation patterns that correctly align with anatomical structure. SSPNet obtains superior performance because it incorporates a dual attention mechanism, which optimizes spatial and channel-wise feature recognition to better detect abnormal regions within MRI images.

The effectiveness of SSPANet is tested through a comparison with Coordinate Attention, SRMNet, GCNet, and SPNet attention modules when applied to VGG16 and ResNet50 base networks, which are shown in Figs. 8 and 9, respectively. Some of these modules succeed at tumor localization, particularly SRMNet; however, they produce broader regions and suffer from noise artifacts. For example, the combination of ResNet50 + SRMNet captures part of the tumor region in glioma utilizing GradCAM++, but it fails to maintain sharp boundary localization and includes irrelevant zones. Similarly, the heatmaps generated by Coordinate Attention and SPNet exhibit lower precision in their spread patterns and show imprecise results among meningioma and pituitary examples. On the other hand, SSPANet creates better-defined and noise-free concentrated attention maps that persist across all backbone methods and tumor types. It proves that the combination of Z-Pool-based Channel Attention along with Strip and Style Pooling-based Spatial Attention makes SSPANet superior to other attention mechanisms because it efficiently extracts discriminative features from local regions. GradCAM++ proves to be the most successful visualization approach alongside SSPANet and other attention-augmented models in all four figures. The heatmaps generated by GradCAM++ are more localized than those produced by GradCAM, which show diffused and imprecise results. EigenGradCAM also shows better visualizations, particularly in the VGG16 model, although it produces minor unwanted background artifacts. SSPANet, together with GradCAM++, resulted in optimal explainability outputs through both VGG16 and ResNet50 models by providing highly accurate detection of tumor regions with correct anatomical alignment.

The explainability evaluations of standard and SSPANet-enhanced models indicate that the proposed SSPANet demonstrates better model

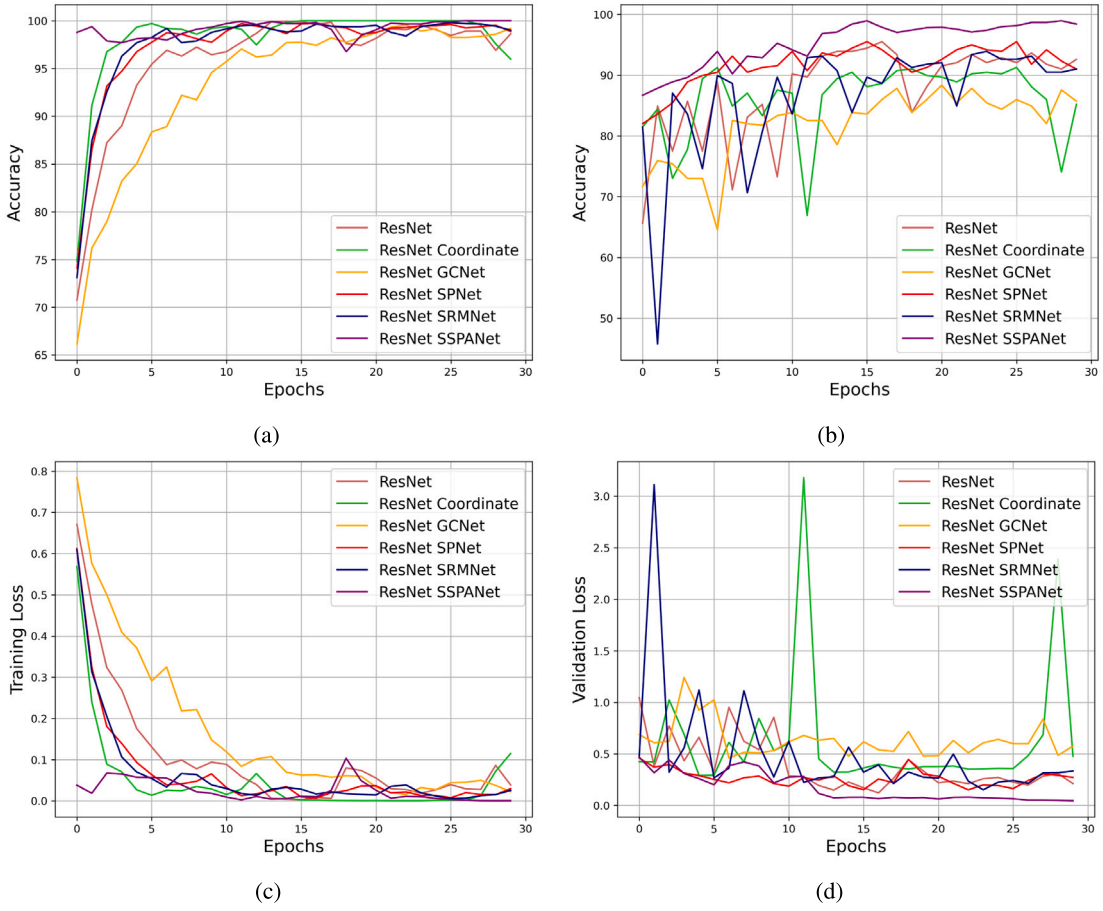

Fig. 5. Accuracy and loss of different attention networks with ResNet50. (a) Training Accuracy and, (b) Validation Accuracy (c) Training Loss (d) Validation Loss.

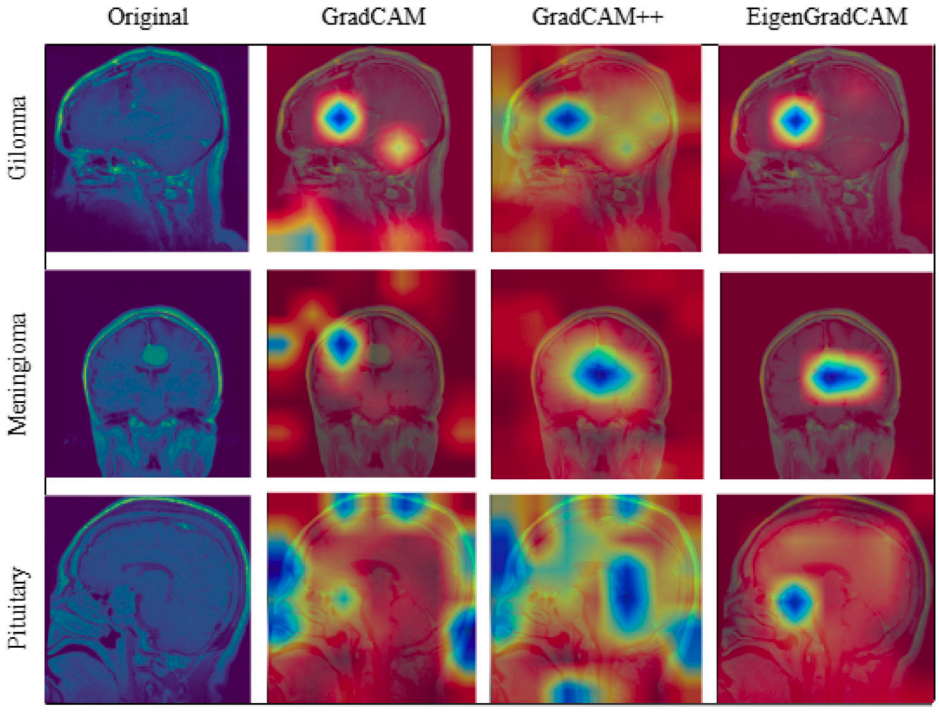

Fig. 6. Explainability result of VGG16 with the proposed SSPANet.

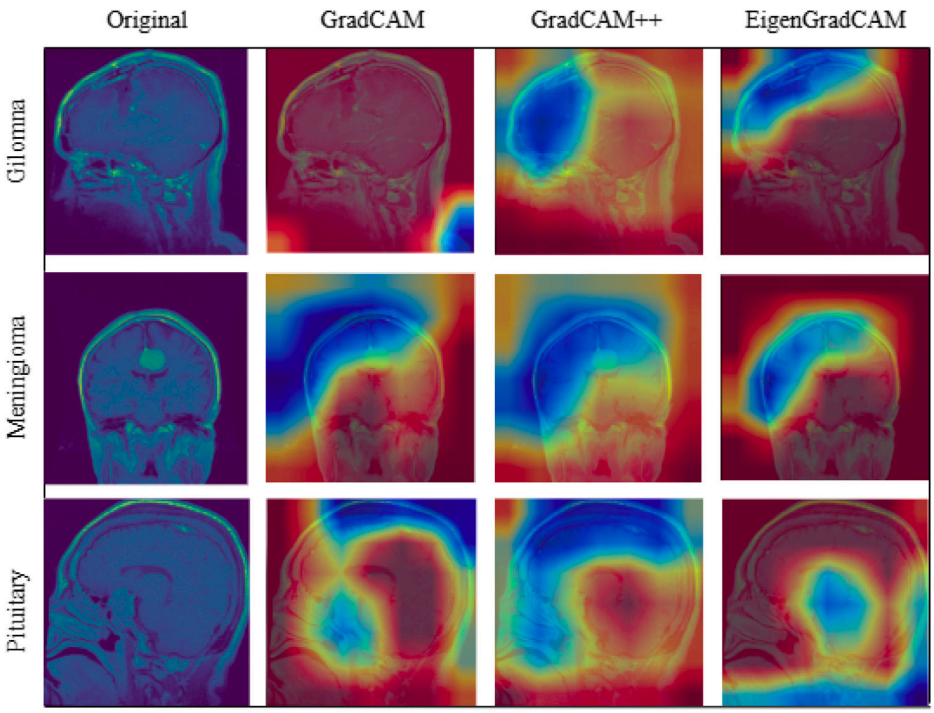

Fig. 7. Explainability result of ResNet50 with the proposed SSPANet.

attention towards tumor areas while lowering nonspecific activation patterns. The dual attention mechanism of SSPANet leads to better feature extraction, together with improved boundary detection, and specialists can easily interpret these visual representations. The reliability and consistency of tumor localization with SSPANet outperform other state-of-the-art attention mechanisms when working with different deep neural network structures and brain tumor categories. The experimental findings demonstrate that SSPANet acts as an effective portable medical imaging enhancement module, positioning it as a promising foundation for AI-based medical diagnosis systems in clinical practice. While our current qualitative analysis emphasizes comparative attention visualization using different GradCAM variants, further exploration on the integration of ground-truth segmentation overlays is needed to further validate alignment between model attention and annotated tumor regions. As the Figshare dataset lacks spatial annotations, quantitative evaluation using metrics like IoU or Pointing Game was not feasible. Instead, we manually validated the GradCAM-based heatmaps against the input images to ensure alignment with visible tumor regions, following practices used in similar non-annotated studies. While ground-truth segmentation overlays can enhance interpretability validation, our study focused on comparative visual analysis of GradCAM, GradCAM++, and EigenGradCAM across tumor types without altering the original layout. The visualizations demonstrate that SSPANet-enhanced models produce more localized and precise activations, particularly with GradCAM++ and EigenGradCAM. We acknowledge the value of incorporating segmentation masks for more detailed assessment and consider this an important direction for future work, especially when pixel-level annotations are available.

4.5. Discussion

This study presents SSPANet, a novel spatial and channel attention mechanism designed to enhance both classification performance and interpretability in BT detection. Integrated with two widely adopted CNN backbones, VGG16 and ResNet50, SSPANet consistently outperforms baseline models and existing attention mechanisms (Coordinate Attention, GCNet, SRMNet, and SPNet) across all key evaluation metrics, including accuracy, precision, recall, F1-score, Cohen's Kappa, and MCC. These improvements are observed across multiple tumor classes: glioma, meningioma, and pituitary.

To further validate the effectiveness of SSPANet, we evaluated its integration with VGG16 and ResNet50 against their respective baselines. As shown in Table 2, SSPANet-VGG16 achieves a 4.5% improvement in classification accuracy over the standard VGG16. Similarly, Table 3 demonstrates a 7.78% accuracy gain when SSPANet is combined with ResNet50. These results confirm that SSPANet enhances the representational capacity of CNNs by introducing more precise and adaptive attention mechanisms. This enables the network to better focus on tumor-relevant regions, particularly in challenging cases such as gliomas, where accurate and localized feature discrimination is essential.

For interpretability, we applied three Grad-CAM variants, such as Grad-CAM, Grad-CAM++, and EigenGrad-CAM, to visualize the model's decision-making process. Grad-CAM provides coarse, class-specific localization maps, while Grad-CAM++ offers finer granularity and handles multiple tumor regions more effectively. EigenGrad-CAM performs eigen decomposition on activation maps, enhancing dominant visual patterns while suppressing noise. Combined with SSPANet, these methods produce sharper and more localized attention heatmaps, which are especially helpful in explaining complex decisions. Notably, ResNet50 + SSPANet combined with Grad-CAM++ yields the clearest activation maps with minimal background interference, facilitating transparent and reliable interpretation.

Our proposed attention-based explainable DL framework demonstrates strong and consistent performance when integrated with benchmark architectures, outperforming existing attention mechanisms across all major evaluation metrics. This consistent performance across both VGG16 and ResNet50 highlights the adaptability and generalizability of SSPANet, showing potential for integration into a wide range of deep learning systems. While the results are promising, further validation on multi-institutional datasets with diverse imaging protocols is essential to ensure generalizability in real-world clinical environments. Furthermore, we acknowledge that rare and ambiguous tumor cases, such as those with low contrast, overlapping tissues, or irregular boundaries, may challenge the model's reliability. Future

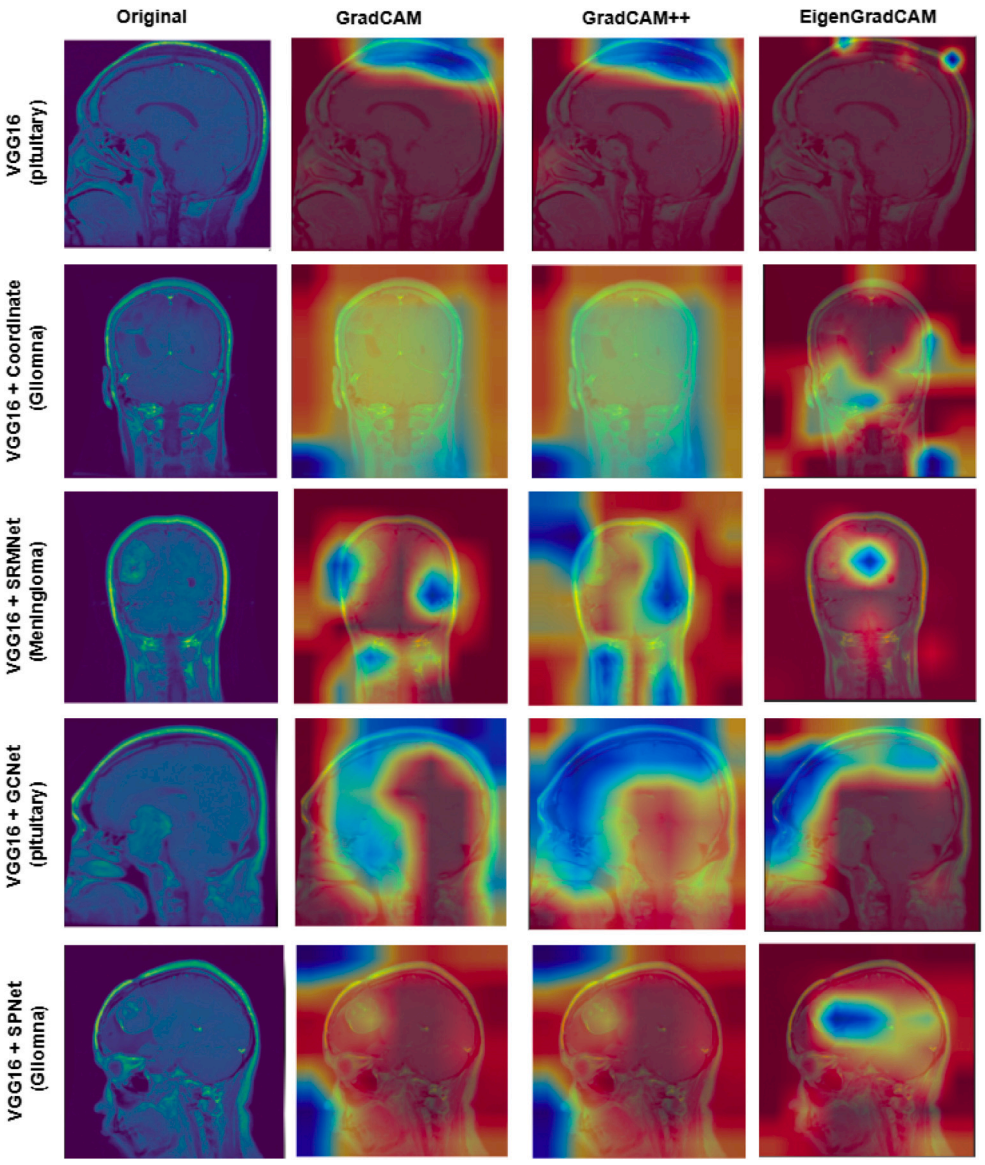

Fig. 8. Explainability result of VGG16 with traditional attentions.

work can incorporate uncertainty quantification, ensemble strategies, or human-in-the-loop feedback to improve decision confidence. From a deployment perspective, optimizing SSPANet for resource-constrained settings through model compression or lightweight architectural design will be crucial for real-time clinical integration, such as in telemedicine or portable diagnostic tools.

All the experimental and explainability results clearly indicate that SSPANet offers a powerful combination of high accuracy and interpretability, contributing to the development of transparent, trustworthy, and clinically viable AI systems for brain tumor classification. This work lays the foundation for further exploration into explainable and robust medical image analysis in practice.

5. Theoretical and computational complexity analysis

In this section, we provide the theoretical analysis to justify the effectiveness and efficiency of the proposed SSPANet attention mechanism. To justify this, we first define some necessary assumptions and then state and prove some supporting lemmas to simplify the proof of the theorems, and finally state and prove the theorems to justify their effectiveness and efficiency.

5.1. Preliminaries and assumptions

Consider $\chi \in \mathbb{R}^{C\times H\times W}$ as the input feature map to the attention mechanism. Then we state three assumptions for signal and noise, attention modulation, and attention alignment, respectively. The assumptions are stated below:

Assumption 1 (*Signal and Noise*)**.** The feature map χ consists of a signal component χ_s and a noise component χ_n, such that $\chi = \chi_s + \chi_n$, with $\mathbb{E}[\chi_s] \gg \mathbb{E}[\chi_n]$.

Intuition: In practical feature maps, informative patterns (signal) dominate over random fluctuations (noise). This assumption ensures that our attention mechanism will have meaningful content to focus on.

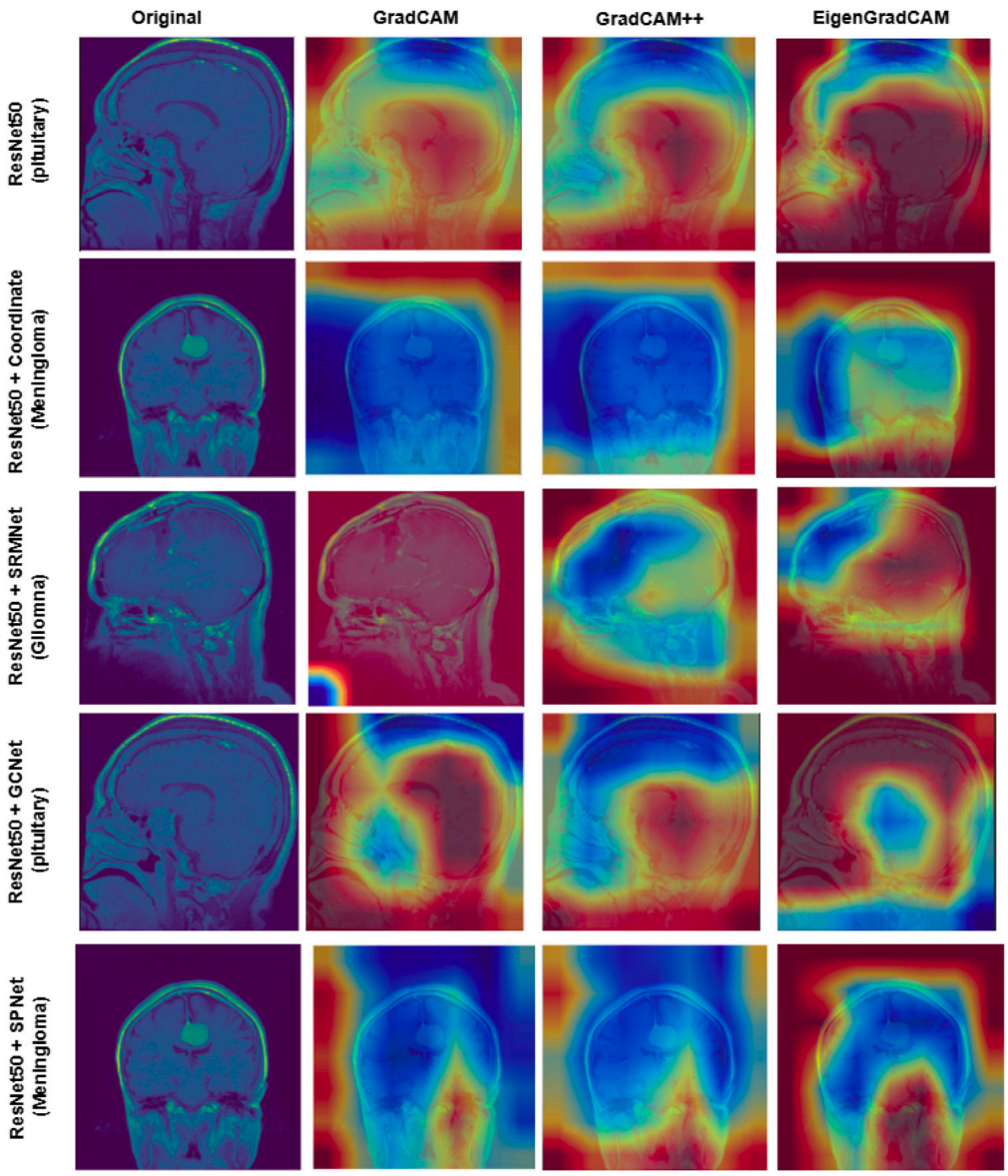

Fig. 9. Explainability result of ResNet50 with traditional attentions.

Assumption 2 (*Attention Modulation*)**.** Attention operates via element-wise modulation as (35).

$$\chi' = M(\chi) \odot \chi = \sigma(M_c(\chi) + M_s(\chi)) \odot (\chi_s + \chi_n). \tag{35}$$

Assumption 3 (*Attention Alignment*)**.** The channel and spatial attention maps are aligned with the signal as in Eq. (36).

$$\mathbb{E}[M_c(\chi_s)] > \mathbb{E}[M_c(\chi_n)], \quad \mathbb{E}[M_s(\chi_s)] > \mathbb{E}[M_s(\chi_n)]. \tag{36}$$

5.2. Lemmas and proof

To simplify the proof of the theorems, we state two important lemmas and prove them separately. The two lemmas are based on the building blocks of the proposed SSPANet, where the first one provides insights into signal-to-noise ratio (SNR) and the second one provides spatial awareness.

Lemma 1. *Z-Pool-Based Channel Attention Improves SNR*

In the proposed SSPANet, Z-Pool-based channel attention improves the SNR in the channel domain more effectively than average or max pooling individually.

Proof of Lemma 1. It is known that traditional SE-Net uses average pooling to compress channel-wise spatial information. SSPANet combines both max pooling and average pooling as Eq. (37).

Intuition: Max pooling captures the most salient (strongest) activation, while average pooling summarizes the overall trend. By combining both, Z-Pool captures diverse statistical information, making it easier for the network to differentiate signal from noise.

$$Z\text{-}pool(\chi) = [\text{MaxPool}(\chi), \text{AvgPool}(\chi)]. \tag{37}$$

This dual-pooling representation improves both salient (maximal) and contextual (average) reactions. The next convolutional layer learns weights to emphasize high-signal activations while suppressing noise. Since the signal χ_s contributes more significantly to both max and average values than noise χ_n, the learned channel weights are more aligned with χ_s than in the single pooling case. Thus,

$$\text{SNR}_{Z\text{-}pool} > \text{SNR}_{\text{AvgPool}}, \quad \text{SNR}_{\text{MaxPool}} \tag{38}$$

This means that the attention mechanism derived from Z-Pool better distinguishes useful information (signal) from irrelevant content (noise), leading to improved channel-wise attention modulation.

Lemma 2. *Strip + Style Pooling Enhances Spatial Awareness*

Compared to traditional square-kernel spatial attention techniques, the spatial attention module in SSPANet enhances spatial context encoding more successfully by utilizing strip pooling and style pooling.

Proof of Lemma 2. Standard spatial attention uses local square kernels (e.g., 7 × 7), which have a limited receptive field. Traditional $k \times k$ kernels (e.g., 7 × 7) can only "see" limited parts of the image at once. But meaningful spatial patterns, such as edges or textures, often span large distances. To capture such patterns, we apply strip pooling that can 'see' across entire rows or columns. SSPANet applies strip pooling as Eq. (39).

$$X_h \in \mathbb{R}^{C\times H\times 1}, \quad X_v \in \mathbb{R}^{C\times 1\times W}. \tag{39}$$

Capturing long-range vertical and horizontal interdependence. Style pooling uses average and standard deviation statistics along each strip to better summarize the global context. Adaptive modulation depending on location and neighborhood context is made possible by passing these statistics through 1D convolutions. The network's effective receptive field is increased by this architecture, which increases its sensitivity to global layout, directional structure, and appearance patterns. Therefore,

$$\mathrm{ERF}_{\mathrm{SSPA}} > \mathrm{ERF}_{\text{standard spatial attention}}. \tag{40}$$

Hence, SSPANet gains a larger effective receptive field (ERF), allowing it to make more spatially-aware decisions when applying attention, especially useful for elongated structures or repetitive textures in input data.

5.3. Theorems with proofs

Theorem 1. *SSPANet Improves Discriminative Feature Representation*

When compared to traditional attention mechanisms, SSPANet improves downstream discriminative performance by better aligning output features with signal components under Assumptions 1–3.

Proof of Theorem 1. *Putting Lemmas 1 and 2 together:* Both channel-wise and spatial attention are more strongly activated by signals than by noise. This means the attention map $M(\chi)$ inherently prioritizes meaningful features, and we have Eq. (41).

$$\begin{aligned}&\mathbb{E}[M_c(\chi_s)] \gg \mathbb{E}[M_c(\chi_n)]\\ &\mathbb{E}[M_s(\chi_s)] \gg \mathbb{E}[M_s(\chi_n)].\end{aligned} \tag{41}$$

Therefore,

$$M(\chi) = \sigma(M_c + M_s). \tag{42}$$

Eq. (42) has higher activation on χ_s and lower activation on χ_n. When this is element-wise multiplied with χ, the noise is suppressed and the signal is preserved as Eq. (43).

$$\begin{aligned}\chi' &= M(\chi) \odot \chi = M(\chi) \odot (\chi_s + \chi_n)\\ &= M(\chi) \odot \chi_s + M(\chi) \odot \chi_n.\end{aligned} \tag{43}$$

Since $M(\chi)$ is more aligned with χ_s, it acts like a filter: boosting useful parts and reducing noisy parts. $M(\chi)$ has low correlation with χ_n, the output χ' approximates an enhanced version of χ_s, improving the feature's discriminative power for downstream tasks (e.g., classification, detection). The final output χ' is therefore a cleaner, more informative version of the original features. Thus:

$$f(\chi') > f(\chi), \quad f(\chi') > f(\chi^{SE}), \quad f(\chi') > f(\chi^{CBAM}). \tag{44}$$

where $f(\cdot)$ denotes task performance.

This results in better performance in tasks like classification or detection because the model learns from clearer, more relevant features. Our experimental results support this theoretical claim.

Theorem 2. *SSPANet has low computational footprint*

The SSPANet architecture provides significant attention-based enhancement with minimal additional computational complexity compared to existing methods.

Proof of Theorem 2. Let C, H, and W denote the channel, height, and width dimensions, respectively. The additional operations in SSPANet include

- Z-Pool: Involves two simple pooling operations (Avg and Max), each $O(HW)$
- Channel Conv: 2D convolution with a small kernel $k \times k$ on 2-channel input → negligible compared to full feature maps
- Strip Pooling: $O(H)$ and $O(W)$ operations over feature columns/rows
- Style Pooling: Global mean and std dev — constant cost $O(C)$ per strip
- 1D Convs: Low-cost, as they operate on reduced spatial dimensions

All steps avoid high-dimensional matrix multiplications (as in non-local attention) or large kernel convolutions. Compared to CBAM, which uses 2 large spatial convs and 2 MLPs, SSPANet substitutes these with lightweight pooling + conv operations. Therefore, the computational complexity:

$$\mathcal{O}_{\mathrm{SSPANet}} < \mathcal{O}_{\mathrm{CBAM}}, \quad \approx \mathcal{O}_{\mathrm{SE}} \tag{45}$$

This makes SSPANet especially appealing for real-time or resource-constrained applications where both accuracy and efficiency are important.

Summary: Through a careful combination of Z-Pool and Strip + Style Pooling, SSPANet creates an attention map that highlights signal and suppresses noise, both in channel and spatial dimensions. The theoretical analysis shows that SSPANet not only improves feature quality but does so with minimal computational cost. These insights are validated by our experimental results.

5.4. Computational complexity analysis

We present the computational footprint of the proposed SSPANet with SE-Net, CBAM, ECA-Net, Coordinate Attention, GCNet, SPNet, and SRMNet attention mechanisms in Table 5. The SE-Net employs global average pooling followed by two fully connected layers for channel attention, resulting in little overhead but lacking spatial awareness. Additionally, CBAM shows more complexity by adding spatial attention through huge 2D convolutions. Conversely, ECA-Net shows efficiency but loses expressive spatial capacity by integrating 1D convolution. Both the Coordinate and SPNet improve directional and long-range context using axis-wise or strip pooling, but ignore style or semantic cues. GCNet performs global context modeling with better performance but contains significant quadratic complexity. SRMNet does not include directed spatial modeling, as it only focuses on style (mean and standard deviation). The proposed SSPANet distinguishes itself by integrating Z-Pool-based Channel Attention with Strip + Style Pooling Spatial Module, resulting in a complexity of $O(C(H + W))$, which is much lower than GCNet and comparable to SPNet and Coordinate Attention. Despite its lightweight design, SSPANet effectively captures channel semantics, spatial structure, and style information all at once, a feature that no other module in the comparison provides. These benefits make it suitable for high-performance and resource-constrained applications.

Table 5
Computational complexity comparison of attention mechanisms, where H, W, and C denote the height, width, and number of input channels of the feature map, r is the reduction ratio, and k is the kernel size.

Module	Main operations	Extra parameter	Complexity	Notes
SE-Net	Global AvgPool → FC → ReLU → FC → Sigmoid	$2C^2/r$	$2C^2/r + C$	Focus on Channel only; lightweight but lacks spatial awareness.
CBAM	SE + Spatial (Avg + Max Pool → 7 × 7 Conv)	$2C^2/r + 7 \times 7$	$\mathcal{O}\left(\frac{2C^2}{r}\right) + \mathcal{O}\left(2HWk^2\right)$	Complexity is higher due to spatial convolution.
ECA-Net	Global AvgPool → 1D Conv	k	$C \cdot k$	Very efficient but less expressive.
Coordinate Attention	AvgPool along H and W → 1D Convs → Fusion	$\approx 2C^2/r$	$\mathcal{O}(C(H+W))$	Efficient but require long-range encoding.
GCNet	Context modeling via attention-weighted sum → 1 × 1 Convs	$\sim C^2$	$\mathcal{O}(HWC^2)$	Global modeling; computationally heavy.
SPNet	Strip pooling over multiple scales → MLP/Conv	Moderate	$\mathcal{O}(C(H+W))$	Directional long-range context.
SRMNet	Style features (mean + std) → FCs → Gate	$\mathcal{O}(C)$	$\mathcal{O}(C)$	Efficient; Style-aware but lacks spatial context.
SSPANet (Ours)	Z-Pool + Conv + BN + Sigmoid + Strip Pooling + Style Pooling + 1D + 1 × 1 Conv → Sigmoid	Low (Conv + 1D)	$\mathcal{O}(C(H+W))$	Combines semantics, spatial structure, and style; Highly efficient.

6. Conclusion and future work

It is crucial to classify or detect brain tumors accurately since this improves patient prognosis and survival rates, allows for early diagnosis, and directs prompt medical intervention. This study focuses on designing a novel attention mechanism aimed at enhancing the performance of existing DL models. Experimental results demonstrate that baseline models struggle to achieve accurate tumor detection due to the complex patterns inherent in MRI images. However, the integration of an attention mechanism significantly improves their capabilities in medical image classification. The proposed attention module, SSPANet, integrates a Z-Pool-based Channel Attention mechanism with a Strip and Style Pooling Spatial Module. This design enables efficient feature extraction, leading to superior performance when applied to both VGG16 and ResNet50 architectures. The high detection accuracy achieved with relatively low computational complexity underscores the efficacy and superiority of SSPANet in brain tumor detection.

In terms of explainability, integrating SSPANet with VGG16 and ResNet50 yields exceptional results. The attention mechanism enhances each model's ability to focus on actual tumor regions within MRI scans. When combined with explainability techniques such as GradCAM, GradCAM++, and EigenGradCAM, the models generate meaningful visual justifications for their predictions, highlighting tumor-affected areas in a clinically interpretable manner. Specifically, EigenGradCAM performs best with VGG16, while GradCAM++ offers superior localization when used with ResNet50. These combinations not only produce low-noise heatmaps but also highlight the most critical regions, thereby demonstrating improved feature extraction and accurate tumor localization. This improvement in interpretability not only strengthens the reliability of AI-assisted diagnosis but also builds trust among medical professionals by offering transparent insights into the model's decision-making process. The synergy between attention-enhanced models and explainability techniques represents a significant step forward in developing robust, trustworthy, and deployable AI systems for medical imaging. Ultimately, these advancements contribute to earlier tumor detection and more effective treatment planning for BT patients.

Further research in BT detection should emphasize the integration of multi-modal data to enhance diagnostic precision and support personalized treatment planning. The development of lightweight, real-time models suitable for deployment on edge devices may improve accessibility in resource-limited settings. Additionally, the adoption of self-supervised and semi-supervised learning approaches can help overcome the challenges posed by limited annotated medical data.

CRediT authorship contribution statement

Md Jahid Hasan: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Mahmudul Hasan:** Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Sumya Akter:** Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Abu Bakar Siddique Mahi:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Md Palash Uddin:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Data curation, Conceptualization.

Ethical approval and consent to participate

This study used publicly available and fully anonymized secondary data obtained from online sources. No direct human participation or identifiable personal information was involved. Therefore, ethical approval and informed consent were not required in accordance with institutional and international guidelines.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to proofread. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The links to the data have been included in the manuscript.

References

[1] Y. Zakeri, B. Karasfi, A. Jalalian, A review of brain tumor segmentation using MRIs from 2019 to 2023 (statistical information, key achievements, and limitations), J. Med. Biol. Eng. 44 (2) (2024) 155–180.

[2] J. Sperber, E. Owolo, N. Abu-Bonsrah, C. Neff, C. Baeta, C. Sun, T. Dalton, D. Sykes, B.L. Bishop, C. Kruchko, et al., Association between urbanicity and outcomes among patients with spinal cord ependymomas in the United States, World Neurosurg. 181 (2024) e107–e116.

[3] K. Agrawal, S. Asthana, D. Kumar, Role of oxidative stress in metabolic reprogramming of brain cancer, Cancers 15 (20) (2023) 4920.

[4] S.K.R. Chinnam, V. Sistla, V.K.K. Kolli, Multimodal attention-gated cascaded U-Net model for automatic brain tumor detection and segmentation, Biomed. Signal Process. Control. 78 (2022) 103907.

[5] E.-S.A. El-Dahshan, H.M. Mohsen, K. Revett, A.-B.M. Salem, Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm, Expert Syst. Appl. 41 (11) (2014) 5526–5545.

[6] S. Anantharajan, S. Gunasekaran, T. Subramanian, R. Venkatesh, MRI brain tumor detection using deep learning and machine learning approaches, Meas.: Sens. 31 (2024) 101026.

[7] S. Saeedi, S. Rezayi, H. Keshavarz, S. R. Niakan Kalhori, MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques, BMC Med. Inform. Decis. Mak. 23 (1) (2023) 16.

[8] C. Qiu, Y. Song, Y. Liu, Y. Zhu, K. Han, V.S. Sheng, Z. Liu, MMMViT: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities, Biomed. Signal Process. Control. 90 (2024) 105827.

[9] Z. Xu, J. Tang, C. Qi, D. Yao, C. Liu, Y. Zhan, T. Lukasiewicz, Cross-domain attention-guided generative data augmentation for medical image analysis with limited data, Comput. Biol. Med. 168 (2024) 107744.

[10] S. Asif, M. Zhao, F. Tang, Y. Zhu, An enhanced deep learning method for multi-class brain tumor classification using deep transfer learning, Multimedia Tools Appl. (2023) 1–28.

[11] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, S. Yin, Deep learning attention mechanism in medical image analysis: Basics and beyonds, Int. J. Netw. Dyn. Intell. (2023) 93–116.

[12] M.I. Nazir, A. Akter, M.A.H. Wadud, M.A. Uddin, Utilizing customized CNN for brain tumor prediction with explainable AI, Heliyon 10 (20) (2024).

[13] T.K. Dutta, D.R. Nayak, Y.-D. Zhang, Arm-net: Attention-guided residual multiscale cnn for multiclass brain tumor classification using mr images, Biomed. Signal Process. Control. 87 (2024) 105421.

[14] K. Md. Hasib, M. Oli Ullah, M. Imran Nazir, A. Akter, M. Saifur Rahman, Icdp: An improved convolutional neural network model to detect pneumonia from chest x-ray images, in: International Conference on Big Data, IoT and Machine Learning, Springer, 2023, pp. 467–479.

[15] A. Azgar, M. Imran Nazir, A. Akter, M. Saddam Hossain, M. Anwar Hussen Wadud, M. Reazul Islam, MNIST handwritten digit recognition using a deep learning-based modified dual input convolutional neural network (DICNN) model, in: International Congress on Information and Communication Technology, Springer, 2024, pp. 563–573.

[16] S. Mandloi, M. Zuber, R.K. Gupta, An explainable brain tumor detection and classification model using deep learning and layer-wise relevance propagation, Multimedia Tools Appl. (2023) 1–31.

[17] R. Haque, M.M. Hassan, A.K. Bairagi, S.M. Shariful Islam, NeuroNet19: an explainable deep neural network model for the classification of brain tumors using magnetic resonance imaging data, Sci. Rep. 14 (1) (2024) 1524.

[18] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, et al., CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation, IEEE Trans. Med. Imaging (2023).

[19] K.R. Pedada, B. Rao, K.K. Patro, J.P. Allam, M.M. Jamjoom, N.A. Samee, A novel approach for brain tumour detection using deep learning based technique, Biomed. Signal Process. Control. 82 (2023) 104549.

[20] M.A. Talukder, M.M. Islam, M.A. Uddin, A. Akhter, M.A.J. Pramanik, S. Aryal, M.A.A. Almoyad, K.F. Hasan, M.A. Moni, An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning, Expert Syst. Appl. (2023) 120534.

[21] D. Rastogi, P. Johri, V. Tiwari, A.A. Elngar, Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework, Biomed. Signal Process. Control. 88 (2024) 105602.

[22] M. Geetha, V. Srinadh, J. Janet, S. Sumathi, Hybrid archimedes sine cosine optimization enabled deep learning for multilevel brain tumor classification using mri images, Biomed. Signal Process. Control. 87 (2024) 105419.

[23] X. Li, X. Fang, G. Yang, S. Su, L. Zhu, Z. Yu, TransU2-Net: an effective medical image segmentation framework based on transformer and U2-Net, IEEE J. Transl. Eng. Heal. Med. (2023).

[24] X. Lin, M. Wang, F. Li, Z. Xu, J. Chen, X. Chen, C. Yuan, S. Wu, Y. Luo, J. Shen, et al., Improving tumor classification by reusing self-predicted segmentation of medical images as guiding knowledge, IEEE J. Biomed. Heal. Inform. (2023).

[25] M.O. Ullah, S.A. Raju, M.I. Nazir, A. Akter, M.S. Rahman, An innovative machine learning pipeline for stroke prediction on imbalanced data, in: 2023 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD, IEEE, 2023, pp. 153–157.

[26] J. Cheng, Brain tumor dataset, 2017, http://dx.doi.org/10.6084/m9.figshare.1512427.v8, Dataset.

[27] S. Mascarenhas, M. Agarwal, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification, in: 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications, CENTCON, Vol. 1, IEEE, 2021, pp. 96–99.

[28] M.I. Nazir, M.O. Ullah, A. Akter, An efficient deep transfer learning based apple leaf disease classification, in: 2023 26th International Conference on Computer and Information Technology, ICCIT, IEEE, 2023, pp. 1–6.

[29] B. Koonce, ResNet 50, in: Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, Springer, 2021, pp. 63–72.

[30] M.O. Ullah, M.I. Nazir, A. Akter, S.A. Raju, M.S.R. Oion, Dry food classification using hybrid deep transfer learning, in: 2023 26th International Conference on Computer and Information Technology, ICCIT, 2023, pp. 1–6, http://dx.doi.org/10.1109/ICCIT60459.2023.10441650.

[31] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[32] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[33] Q. Hou, L. Zhang, M.-M. Cheng, J. Feng, Strip pooling: Rethinking spatial pooling for scene parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4003–4012.

[34] H. Lee, H.-E. Kim, H. Nam, Srm: A style-based recalibration module for convolutional neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1854–1862.

[35] C.C. Ukwuoma, D. Cai, E.O. Eziefuna, A. Oluwasanmi, S.F. Abdi, G.W. Muoka, D. Thomas, K. Sarpong, Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI–LIME & SHAP, Biomed. Signal Process. Control. 100 (2025) 107014.

[36] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[37] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, pp. 839–847.

[38] M.B. Muhammad, M. Yeasin, Eigen-cam: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–7.