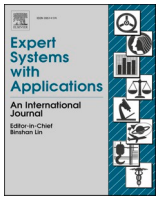
Expert Systems With Applications 241 (2024) 122672

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Vision transformer based classification of gliomas from histopathological images

Evgin Goceri

Department of Biomedical Engineering, Engineering Faculty, Akdeniz University, Turkey

ARTICLE INFO

Keywords:
Brain tumor
Classification
Deep learning
Digital pathology
Glioma
Transformer

ABSTRACT

Early and accurate detection and classification of glioma types is of paramount importance in determining treatment planning and increasing the survival rate of patients. At present, diagnosis in neuropathology is based on molecular and histological characteristic information provided with microscopic visual examinations of biopsies. However, the traditional method is not only laborious, and time-consuming but also needs experience. Furthermore, the subjective diagnosis causes inter-/intra variability and late or inaccurate diagnosis. To overcome those issues by automated methods, Convolutional Neural Networks (CNNs) and, more recently, transformer-based models have been used. However, they have their own drawbacks. For instance, CNNs ignore global information by focusing on pixel-wise information, although they are good at the extraction of local characteristic features using several convolution and pooling layers. Vision transformers are problematic in the extraction of details and local features, although they are good in the extraction of global features using global receptive fields in the early layers. Therefore, in this work, their advantages have been utilized in designing a new architecture to classify gliomas. Obtaining high performance from the proposed architecture has been achieved by (i) using a combined version of CNN and transformer stages, and (ii) integrating effectively designed feature-combining and smart-joining modules appropriately. Experiments have indicated the effectiveness of the proposed approach in classifying four glioma subtypes from histopathological images in terms of several evaluation metrics (i.e., accuracy (96.75%), recall (97.00%), precision (96.75%), F1-score (96.80%)). Comparative evaluations of the performances of the state-of-the-art techniques have shown better capability of the proposed approach.

1. Introduction

Gliomas are considered as potentially fatal brain cancers among various cancer types. They arise mainly in consequence of the cancerization of oligodendrocytes, astrocytes, and glial cells. The main symptoms of glioma patients are cognitive and neurological dysfunction, increasing intracranial pressure, and seizures. They can be caused by a variety of reasons (e.g. family history, radiation exposure, and age) (Zhang et al., 2022). Gliomas are one of the most widespread primary tumors comprising almost 80 % of malignant brain tumors (Sung et al., 2021). Their annual incidence is six cases per hundred thousand people worldwide and is approximately 1.6 times more common in men than in women (Ostrom et al., 2019). While 5-year survival rates reach eighty percent in patients having low-grade glioma (e.g. oligodendroglioma), this rate is below five percent in patients having high-grade glioma (e.g. glioblastoma) (Komori T: The, 2021). This means that patient survival mainly depends on glioma subtypes. Therefore, accurate and early detection of glioma subtypes is of utmost importance in diagnosis and treatment planning.

Today's gold standard approach in diagnosis made by neuropathologists is to visually examine the genetic/molecular and/or morphological properties of tissues obtained by biopsy under a microscope. Information of the morphological properties is obtained after staining histopathological sections on glass slides. In the staining process, generally, Hematoxylin & Eosin (H&E) biomarkers are used to stain nuclei with blue/purple and to stain connective tissues and cytoplasm with pink/red color, respectively. Determination of a tumor type is based on a combination of morphological information provided from the stained slides, and molecular and immune-histochemical information (Komori T: The, 2021; Perry and Wesseling, 2016). Example images showing four sub-types of glioma are presented in Fig. 1 (National Cancer Institute, 2023).

Diagnosis with the gold standard procedure causes significant intra-observer and inter-observer variabilities. Because it needs experience

E-mail address: evgin@akdeniz.edu.tr.

https://doi.org/10.1016/j.eswa.2023.122672
Received 15 October 2023; Received in revised form 9 November 2023; Accepted 16 November 2023
Available online 25 November 2023

due to large heterogeneities within tumors and morphological variations. Also, it is laborious and time-consuming due to visual examinations of fine and coarse resolutions of images including tissue samples with large volumes. Additionally, there can be in-consistencies even among experienced pathologists on the same tissue sample because of different perceptions and biases (Van den Bent, 2010; Sharma et al., 2015). Therefore, automated methods based on quantitative analyses of histopathological images with advanced computer algorithms are needed for glioma classifications to (i) eliminate the inter-/intra-observer variabilities caused by subjectivity, (ii) reduce pathologists' workloads, (iii) improve diagnostic accuracy, and (iv) provide fast diagnosis.

Recent advances in computer vision algorithms and the growth of digital pathology have increased interest in applying deep learning-based techniques (Liu et al., 2023; Wen et al., 2023; Wu and Moeckel, 2023). Particularly, Convolutional Neural Networks (CNNs) have been preferred in digital pathology (Liu et al., 2023; Wen et al., 2023; Wu and Moeckel, 2023). More recently, vision transformer models have been used in computational histopathology (e.g., image classification tasks) (Xu et al., 2023, 2023; Lan et al., 2023; Atabansi et al., 2023). In vision information transformers, images are partitioned into patches and sent to a transformer network in the form of a linear embedding series of the patches. Leveraging self-attention mechanisms, the transformer models have been shown to outperform CNN structures (Dosovitskiy et al., 2021). Similarly, the higher performance of a vision transformer compared to CNNs has been indicated in a current investigation on the classification of brain tumors from histopathological images (Li et al., 2023).

It has been observed in a recent work that the combination of a vision transformer with a CNN provides better accuracy in comparison with the usage of only CNNs or only a vision transformer (Maurício et al., 2023). Because, vision transformers focus on patch-wise information and can extract global features while CNNs focus on pixel-wise information and can extract local features with the help of pooling and convolution layers. Therefore, in the proposed approach, a combination of them has been applied to achieve the classification of gliomas into four sub-classes (Fig. 1) with high performance.

The proposed architecture has been designed so that the transformer and feature extraction stages are complementary to each other. Also, feature-combining and smart-joining modules have been integrated into the architecture to provide conversions between patches and feature maps and to merge features efficiently. Both global and local characteristics information have been retained and joined intelligently to achieve classifications with high performance. The major contributions of this work include:

i. Introducing a new hybrid architecture utilizing the advantages of CNNs and vision transformers.
ii. Implementation and testing of the proposed method for multi-class classification of gliomas
iii. Application of the current classifiers used for glioma classifications with the same data sets.
iv. Performance comparisons of the methods applied in this work using the same evaluation metrics.

The proposed method can assist pathologists in decision-making by supporting examinations. It can increase the objectivity and diagnostic accuracy. Also, it can reduce pathologists' workload and allow them to spend more time for other complex processes. Therefore, the proposed hybrid approach is promising to overcome the issues caused by the traditional method based on microscopic examinations of glass slides and to replace it.

The remaining sections has been organized as follows. Related works have been given in Section 2. The data sets used in this study and the proposed method have been explained in Section 3 and Section 4, respectively. Discussions and conclusions have been presented in Section 5 and Section 6, respectively.

2. Related works

There are many works in the literature about classification of glioma and its sub-types from magnetic resonance images (Kalaroopan and Lasocki, 2023; Younis et al., 2023; Zhang et al., 2023; Hafeez et al., 2023; Sun et al., 2023; Cluceru et al., 2022). Also, there are many works on the grading of gliomas using transfer learning or transformers from magnetic resonance images (Pitarch et al., 2023; Wu et al., 2023; Khorasani and Tavakoli, 2023; Gilanie et al., 2023). Mostly, those works indicate the high performance of the methods in the differentiation of high- and low-grade gliomas, while the classifications of grades two, three, and four are still a challenging issue. However, there are less works, which have been performed with deep networks recently, on classification of glioma sub-types from pathological images. Because, advances in scanning technologies that convert glass slides into images over the past three years have enabled the scanning of large numbers of slides and encouraged especially the use of deep learning-based methods to help pathologists make early and accurate diagnoses (Shafi and Parwani, 2023).

The methods proposed for classification of glioma sub-types from pathological images and significant information of those methods have been presented in Table 1.

The CNN-based methods in the literature can accomplish glioma and its sub-types classifications from histopathological images automatically. However, each of them has its drawbacks or limitations (such as binary classification of gliomas as low-/high-grade glioma). In a more recent work, the deep network that has been constructed to utilize advantages of both convolutional layers and self-attention layers from transformers has been proposed to classify gliomas into four classes (glioblastoma, oligodendroglioma, astrocytoma and low-grade astrocytoma) (Wang et al., 2023). Experiments indicate that the network produces results with accuracy, sensitivity, and specificity of 77.3 %, 76.0 %, and 86.6 %, respectively. However, its robustness should be

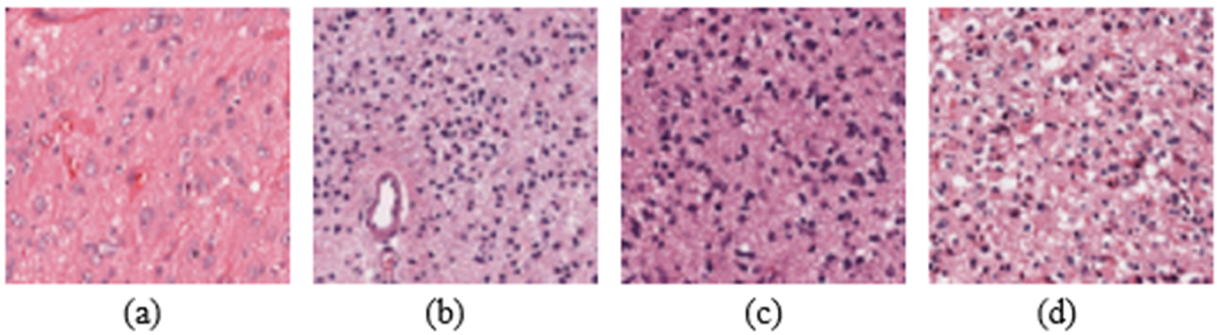


Fig. 1. Glioma sub-types: Glioblastoma (GBM)(a), oligoastrocytomas (b), astrocytomas (c), oligodendrogliomas (d) (National Cancer Institute, 2023).

Table 1
Classifications of glioma types from histopathology images with CNN models.

Reference	Method	Advantage	Disadvantage	Result
(Jose et al., 2023)	Glioma images are classified into three classes (astrocytoma, oligodendroglioma, and glioblastoma) with two networks (ResNet50 and VGG19)	The ResNet50 model benefits from shortcut connections to learn the residuals between inputs and outputs	The VGG19 and ResNet50 models are not effective to detect and extract global feature information	The ResNet50 model produces better results in terms of accuracy (86.10 %) than the VGG19 model
(Chitnis et al., 2023)	Classification of three types of glioma (astrocytoma, oligodendroglioma, and glioblastoma) using two models (ResNet50 and DenseNet121)	The dense connections in the DenseNet121 enhance the classification, the ResNet50 benefits from shortcut connections	Although local features can be captured effectively, extractions of global features are not effective	The accuracy provided from the DenseNet121 (88.24 %) is higher than the accuracy from the ResNet50 (83.67 %)
(Prathaban et al., 2023)	A convolutional network activated by sigmoid function to classify pathological images showing diffuse gliomas (into three classes: cellular tumor, normal brain, and infiltrating edge)	The multi-layer structure with sigmoid function is promising in categorizing cellular tumors, healthy brain tissues, and infiltrating edges	The reliability of the model in the classification of other glioma types is unclear, it should be tested with the images showing other gliomas	The accuracy is 91.70 %, 97.04 %, and 91.62 % in the classification of cellular tumors, healthy brain tissues, and infiltrating edges, respectively
(Pei et al., 2021)	A CNN model, and fusion of cellularity features with molecular features to classify gliomas into two classes as low- and high-grade glioma	The method uses the features that indicate cellularity, which helps to improve pattern recognition from the images	The performance of the method has not been tested with other glioma types, so its reliability is unclear	The accuracies for high-grade glioma versus low-grade glioma are 93.81 % and 73.95 %, respectively
(Jin et al., 2021)	A CNN based on a DenseNet backbone to classify gliomas as oligodendroglioma, glioblastoma, anaplastic oligodendroglioma, anaplastic astrocytoma, astrocytoma	The dense connections in the proposed deep network structure enable efficient extraction of local feature information	The ReLU can cause dying neurons, also randomly changing the colors, contrast, and brightness may cause a loss of information	The deep network model can learn the features and performs the classification of glioma sub-types with 87.5 % accuracy

evaluated since the activation function used in the graph convolutional layers may cause dying neurons and low performance in the classifications. Therefore, an automated method is still needed to achieve multi-class classification of glioma sub-types with high performance. For this purpose, a new hybrid network architecture has been designed and implemented in this study.

3. Data sets

In this work, whole slide images stained with H&E stains and provided from the publicly accessible database called The Cancer Genome Atlas (TCGA) (National Cancer Institute, 2023) have been used to construct data sets. The TCGA is an international multi-centered project aimed at comprehensively analyzing multiple aspects of cancer types (National Cancer Institute, 2023). For our experimental works, a total of 2633 images taken from 926 cases have been used. They include these four types of gliomas: GBM (471), astrocytomas (168), oligodendrogliomas (173), and oligoastrocytomas (114). Those images have been used to construct training (2087 whole slide images of 738 cases), validation (282 whole slide images of 94 cases) and testing (264 whole slide images of 94 cases) datasets randomly. The original images are with 20x resolution and have been cropped into patches with the size of 224 × 224 pixels. Any other preprocessing has not been applied.

4. The proposed method

The advantages of the deep convolutional network structures and vision transformers are combined in the proposed method. Initially, maximum pooling has been applied to preserve the most significant features as well as point out local features. Then, feature maps have been passed through to the CNN (where cascaded convolutions have been applied to capture further spatial information) and transformer path (where global features have been obtained by using stacked self-attention mechanisms).

The functions of the vision transformer and CNN complement each other in the proposed hybrid architecture (Fig. 2). Following each CNN step, feature maps having intensive local features are passed through the transformer path. In the feature-combining module, the features coming from 2 sources are merged, and feature maps are transformed into the structures known as patch embeddings before being sent into the subsequent transformer. By using the self-attention mechanisms in the transformer blocks, non-local dependencies are modeled. After being

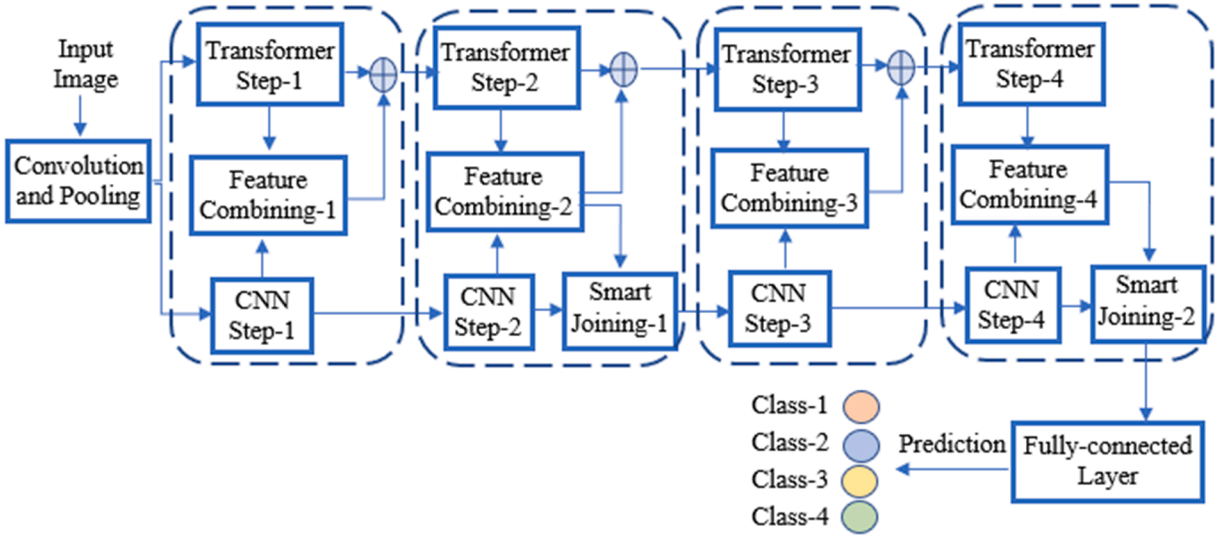


Fig. 2. The proposed hybrid architecture.

separated from the transformer steps, the patch embeddings provided from the transformer are sent backward into the CNN feature extractor via the feature-combining module. Then, both global and local feature information that is coming from the transformer and the CNN, respectively, are combined in the smart-joining module in an intelligent way by choosing the most useful features for combining. To obtain the prediction, the output of the 2nd smart-joining module is passed into a classifier.

4.1. Feature extraction with CNN

CNNs can obtain local feature information in a hierarchically through convolution operations and keep local 2cues as feature maps (Peng et al., 2021). In the proposed method, an efficient CNN model known as DenseNet121 (Huang et al., 2017) has been used. Because, in this architecture, every layer's input includes all outputs coming from the previous layers. Direct access from each layer to the gradients coming from the loss functions can be performed with the dense connection which makes easy spreading of the features over the network. In each step of the DenseNet121 architecture, whose details are presented in Table 2, a stack of convolution layers is used. Following each 3×3 convolution layer, there exists a transition layer including an average pooling layer, a convolution, and also a batch normalization layer.

4.2. Global guidance of features with transformers

Global guidance of the features that are provided from the CNN structures is performed in the transformer pipeline. Before the 1st transformer step, the tensor is projected into a space with high-dimensional and including 768 channels by using convolution operations. By this way, local information can be used from convolutions by the transformer. This is also convenient with conclusion that additions of locality into transformers' early layers enhance feature representation (Dai et al., 2021; Aladhadh et al., 2023). In every transformer step, flattening of the feature maps (obtained by the CNN and have dimensions $B \times C \times H \times W$, where W, H, C, and B terms denote the width, height, number of channels, and batch size, respectively), to patches with dimensions of $B \times (H \times W) \times C$ is performed. Because of this, each pixel is like an individual patch embedding. The flattening process is applied to generate the path sequence, which is then used for projection to the dimension of the transformer. A class token, which can be trained in the training stage to get class-specific information, is linked to these patches. Unlike the conventional vision transformer model (Dosovitskiy et al., 2021), where the trainable class token is typically used for final prediction, in the proposed model, the class token is used to combine with the feature maps and to get global features.

In the proposed method, there exist 4 transformer structures including normalization, multi-head self-attention (MSA) and multi-layered perceptron (MLP). A transformer is represented in Fig. 3.

Table 2
Step, operation, and output size information for the DenseNet121 model.

Step	Operation	Output Size
Convolution	7×7 convolution, stride 2	128×128
Pooling	Maximum pooling, stride 2	64×64
Step-1	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 6$	32×32
Step-2	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 12$	16×16
Step-3	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 64$	8×8
Step-4	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 48$	8×8
Pooling	Global average pooling	1×1
Fully convolution	1024×(number of class)	1×1

In the MSA block, these three matrices are created from the $B \times H \times W \times C$ sized input: Query matrix (R), value matrix (U), and key matrix (L). With these matrices and the softmax function (ψ), the self-attention mechanism is defined by:

$$Attention(R, L, U) = \psi\left(RL^T\right)/\sqrt{D_{Value}}\Big)U \tag{1}$$

The query, value and key matrices are $R \in B \times H \times W \times D_{query}$, $U \in B \times H \times W \times D_{value}$, $L \in B \times H \times W \times D_{query}$, respectively, where the terms C, D_{value}, and D_{query} refer to the sizes of three matrices (i.e., input, value, and query matrix, respectively).

The outputs of the MSA block are passed into the MLP structure consisting of 2 fully-connected layers, which are divided by the activation function known as Gaussian error linear unit. Up-projections of patch embeddings to 3072 dimensions are performed in the 1st fully-connected layer followed by down-projections (which are made to reduce the dimensions to 768) in the 2nd fully-connected layer. The operation in a transformer structure is written using the layer normalization function δ by:

$$T_{MSA\_Output} = MSA\left(\delta\left(T_{Input}\right)\right) + T_{Input} \tag{2}$$

$$T_{Output} = MLP\left(\delta\left(T_{MSA\_Output}\right)\right) + T_{MSA\_Output} \tag{3}$$

In (2) and (3), the terms T_{MSA_Output} and T_{Output} refer to the output of the MSA and transformer structures, respectively, while the term T_{Output} refers to the input for the transformer.

4.3. Feature-Combining

Transformer patch embeddings and CNN feature maps are not consistent in terms of their sizes. This can cause a loss of information if their conversions are performed rigidly. In the feature-combining step, the misalignments between the global and local features coming from the transformer and CNN parts, respectively, are reduced using the input $F_{InputFromCNN}$ from the CNN and $F_{fromTransformer}$ from the transformer part (Fig. 4). The sizes of the inputs $F_{InputFromTransformer}$ and $F_{InputFomCNN}$ are $B \times (1 + H \times W) \times C$ and $B \times C \times H \times W$, respectively. Here 1 refers to the class token, the terms C, B, W, and H correspond to the number of channels, the batch size, and feature maps' width and height values, respectively.

By using global pooling, class-sensitivity of the inputs can be increased, and informative channels can be chosen. Therefore, a global pooling (with 1×1 convolution, Leaky Rectified Linear Unit (LeakyReLU), and batch normalization) is applied on $F_{fromCNN}$ before being converted to patch embeddings, and the channel number of $F_{InputFromCNN}$ is set to 768. Afterwards, down-sampling of the pooled $F_{InputFromCNN}$ is provided by an average pooling since this pooling operation helps to spread the global information by computing the mean value of neighboring pixel values. Because of this, the CNN inputs and the transformer features become more aligned. Following the average pooling (applied with stride 2), feature maps re-shaped to patch embeddings $F_{PatchFromCNN}$. To generate $F_{OutputFromTransformer}$, $F_{InputFromTransformer}$ and $F_{PatchFromCNN}$ are merged.

In the proposed architecture (Fig. 2), four feature-combining modules are used. The output of the former transformer step is summed with each $F_{OutputFromTransformer}$ of a feature-combining module, and then their summation is used as input for the following step. To obtain $F_{OutputFromTransformer}$, $F_{PatchFromCNN}$ is connected to the class token. Apart from the class token, patch embeddings of $F_{InputFromTransformer}$ are discarded in the feature-combining process (Fig. 4). The statistical properties of other patches that are provided from the transformers' inputs are inherited by the class token, which includes class-specific information. The $F_{PatchFromCNN}$ is rich in terms of local features and therefore it guides from CNNs to transformers. After $F_{OutputFromTransformer}$ is used as

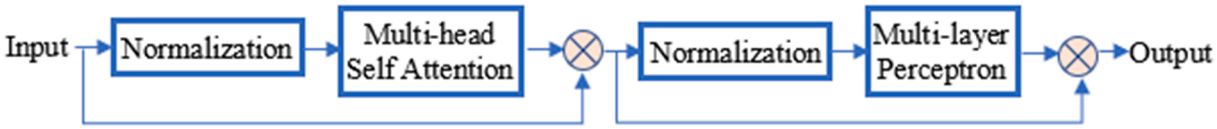


Fig. 3. Transformer block.

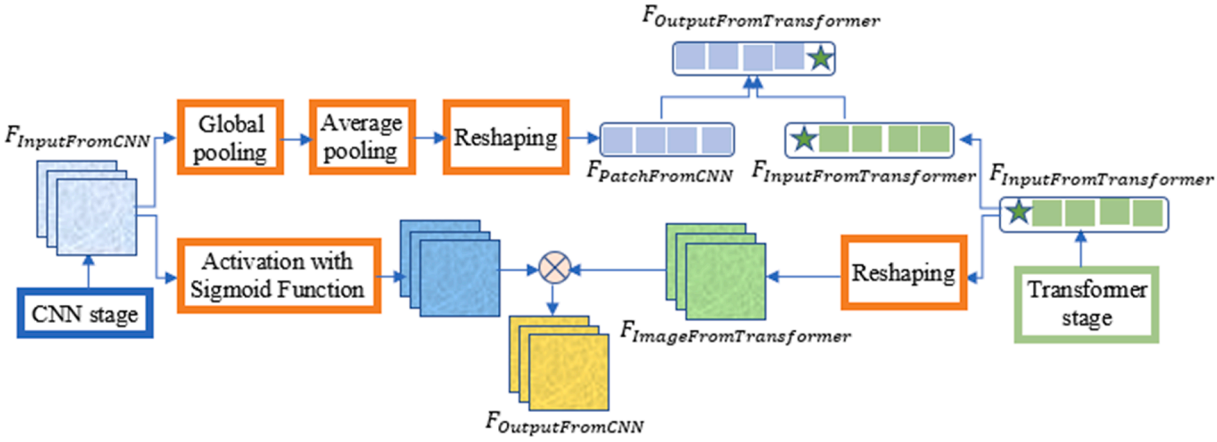


Fig. 4. Feature-combining module.

input to the subsequent transformer, the MSA explores global feature information by using the local feature information from the CNN part to improve the ability of the feature representation.

For the CNN feature-extraction, only the 2nd and 4th feature-combining modules have the outputs $F_{OutputForCNN}$, which are inputs for the 1st and 2nd smart joining modules (Fig. 2). $F_{InputFromTransformer}$ is scaled in the feature-combining module to produce $F_{ImageFromTransformer}$, which is rich in terms of global features. The output $F_{OutputForCNN}$ is obtained using element-wise multiplication ($\otimes$) and sigmoid function (S) by:

$$F_{OutputForCNN} = S\left(F_{InputFromCNN}\right) \otimes F_{ImageFromTransformer} \quad (4)$$

In the multiplication process in (4), $S\left(F_{InputFromCNN}\right)$ behaves like a mask for appropriate spatial support to $F_{ImageFromTransformer}$. By this multiplication process, $F_{OutputForCNN}$ gets the global features from $F_{ImageFromTransformer}$ by inheritance.

4.4. Smart-Joining

In the smart joining process, the transformers' patch embeddings and the feature maps from the CNNs are joined (merged) by capturing and choosing the most valuable data to use in the prediction stage. The smart joining process (Fig. 5) produces an output (I_{Output}) using 2 inputs that are the image coming from the CNN step ($I_{fromCNNstep}$) and the image coming from the feature-combining step ($I_{fromFeatureCombining}$).

There exist a lot of global features in the $I_{fromFeatureCombining}$ that are the same as the $I_{OutputfromCNNstep}$. To detect the most valuable ones among those global features in the $I_{fromFeatureCombining}$ and to join them, relation-

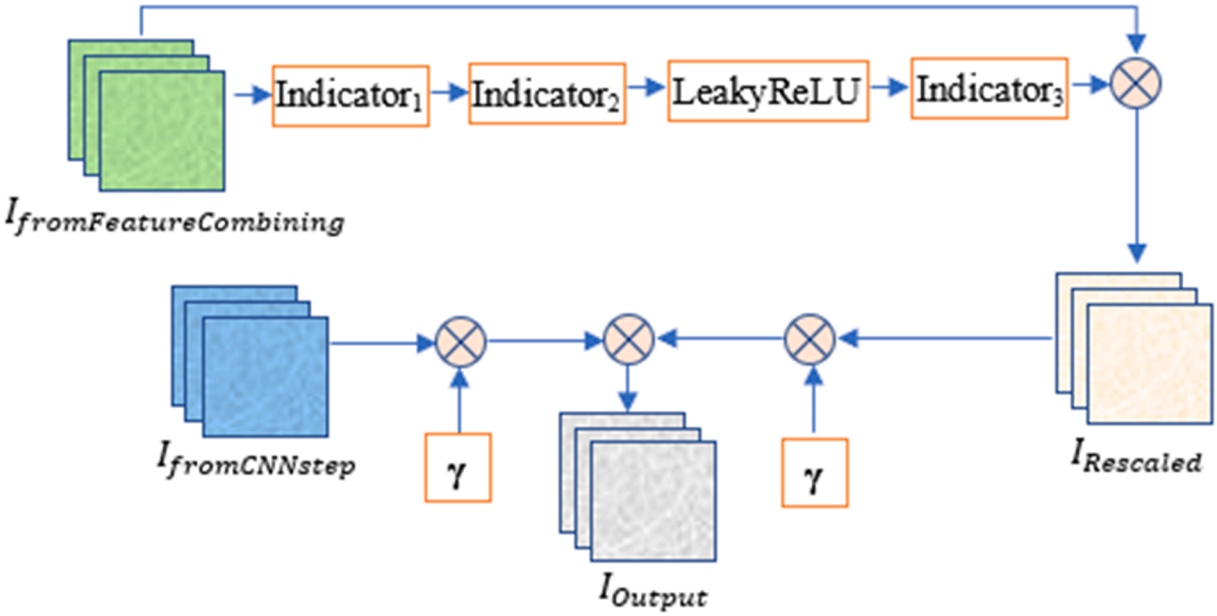


Fig. 5. Smart-joining module.

ships, which are set up between the channels of the $I_{fromFeatureCombining}$, are used.

In Fig. 5, the relationships are shown with indicator $Indicator_3$ having dimension $B \times C \times 1 \times 1$, here B denotes the batch size while the term C denotes the number of channels of $I_{fromFeatureCombining}$. The indicator is obtained with a two-step iterative process. In the first step, a temporary indicator $Indicator_1$ is obtained with global average pooling by using $I_{fromFeatureCombining}$ to merge its spatial features. The pooling operation is useful to generate an indicator since it increases the power of representing global features. The first temporary indicator is defined by:

$$Indicator_1^k = \left(\sum_{i=1}^{H} \sum_{j=1}^{W} \left(I_{fromFeatureCombining} \right)^k (i,j) \right) \Big/ (H/W) \quad (5)$$

In (5), H and W denote the feature maps' height and width properties, $Indicator_1^k$ refers to kth element of $Indicator_1$, and the term $\left(I_{fromFeatureCombining}\right)^k(i,j)$ corresponds to the pixels in the kth channel of $I_{fromFeatureCombining}$. In the second step, compactness of the $Indicator_1$ is provided and another temporary indicator $Indicator_2$ is obtained with $Indicator_2 = L^1_{FullConnected}(Indicator_1)$ by using a full-connected layer $L^1_{FullConnected}$. Here, $Indicator_2 \in B \times (C/\alpha) \times 1 \times 1$, the term $\alpha = 16$ refers to the compact-ratio whose value has been found experimentally in this work. The second indicator is passed to an activation layer (in which LeakyReLU is used). Then, the indicator $Indicator_3$ is obtained with $Indicator_3 = S(L^2_{FullConnected}(Indicator_2))$, where S refers to the sigmoid function, by using another full-connected layer $L^2_{FullConnected}$ which provides restoring the channel numbers. It should be noted here that there exists only 1 pixel on every channel of the $Indicator_3$. Each pixel's value denotes how important the corresponding channel of $I_{fromFeatureCombining}$. Informative channels are highlighted while less informative channels are suppressed by multiplying $I_{fromFeatureCombining}$ with $Indicator_3$.

In the training process, the values in the $Indicator_3$ are updated and the $Indicator_3$ can learn the relationships among the channels of $I_{fromFeatureCombining}$. An intermediate image (represented by $I_{Rescaled}$ in Fig. 5), a re-scaled version of $I_{fromFeatureCombining}$ with $Indicator_3$, is obtained as follows:

$$I_{Rescaled} = I_{fromFeatureCombining} \bullet Indicator_3 \quad (6)$$

The features (that are global features coming from $I_{fromFeatureCombining}$) in $I_{Rescaled}$ are aggregated (added) with the features (that are local features extracted with CNN) in $I_{fromCNNstep}$ to improve the representation ability and accuracy of the proposed method. Their simple aggregations may cause a loss of global or local feature information since their characteristic properties (variance and mean values) can be very different in each channel. Therefore, it is a weighted aggregation, where contributions of these features are determined by weighting parameters that have initial values of 1 and are updated automatically during training after each epoch. The values of the weighting parameters are sequences of the real numbers whose length is identical to the number of channels of $I_{fromCNNstep}$. The summation is performed intelligently thanks to the trainable sequences and is defined with the weighting parameters (λ and γ) by:

$$\left(I_{Output}\right)^i = \lambda^i \bullet \left(I_{fromCNNstep}\right)^i + \gamma^i \bullet \left(I_{Rescalied}\right)^i \quad (7)$$

In (7), the terms $\left(I_{fromCNNstep}\right)^i$ and $\left(I_{Rescalied}\right)^i$ correspond to the ith feature channel of the images $I_{fromCNNstep}$ and $I_{Rescalied}$, respectively, and the term $\left(I_{Output}\right)^i$ refers to the ith channel's feature map of the I_{Output}.

The weighted addition improves the feature representation ability of the proposed architecture. Because, on every feature map, the proportion of global and local features can be obtained and used in the addition. This provides simultaneous retaining of those features. By this way, the local feature information indicating details and coming from the CNN can be enhanced by the global feature information coming from transformers.

5. Results

The performance of the proposed classifier has been evaluated by commonly used evaluation metrics, i.e., accuracy, F1-score, recall, and precision, which can be formulated by:

$$Accuracy = (TN + TP)/(TP + FN + FP + TN) \quad (8)$$

$$Recall = TP/(FN + TP) \quad (9)$$

$$Precision = TP/(FP + TP) \quad (10)$$

$$F1\text{-}score = 2TP/(2TP + FN + FP) 11)$$

The meanings of the terms used in (8)-(11) can be interpreted for a multi-class classification as follows, for example, for GBM:

- TP: The term means True-Positive and denotes the number of images classified as GBM that are GBM.
- FP: The term means False-Positive and denotes the number of images classified as GBM that are not GBM.
- FN: The term means False-Negative and denotes the number of images showing GBM and classified as another disorder.
- TN: The term means True-Negative and denotes the number of images showing not GBM and classified as another disorder.

In addition, comparisons of the performances of the state-of-the-art approaches have been performed using the same metrics. Also, comparisons with the performances of commonly used CNN models (i.e., ResNet50, ResNet101, DenseNet121) have been performed. The LeakyReLU activation function has been used in all of them. The results obtained from all those methods have been presented in Table 3.

For fair comparisons, the methods in Table 3 have been trained with the same datasets that include whole slide images showing four types of gliomas (i.e., GBM, astrocytomas, oligodendrogliomas, oligoastrocytomas) (Section 3). Also, the same validation and testing sets have been used for them. All datasets have been constructed with 224x224 pixel images at 20x resolution.

Also, evaluation metrics have been computed for each class separately to see the performance of the proposed approach for each glioma type separately. The results have been given in Table 4. Also, a confusion matrix (with shorted names for oligoastrocytomas (OA), oligodendrogliomas (OD) and astrocytomas (A)) has been shown in Fig. 6.

In this work, optimization has been provided by stochastic gradient-descent algorithm and a constant learning rate (1×10^{-4}) has been used during the training stage. Cross-entropy loss function has been applied and 200 has been used for the value of the maximum number of epochs parameter.

The proposed method has been applied by using Pytorch framework. Experimental works have been performed on an NVIDIA RTX 3090 GPU system.

Table 3
Performance of the glioma classification methods in percentage.

Method	Accuracy	Precision	Recall	F1-Score
ResNet50	78.18	79.49	78.18	78.14
ResNet101	81.02	78.34	81.02	77.22
DenseNet121	84.54	85.20	84.54	84.59
(Prathaban et al., 2023)	86.22	93.82	86.22	89.16
(Pei et al., 2021)	88.22	79.12	88.22	82.38
(Jin et al., 2021)	90.37	90.44	90.37	90.36
(Wang et al., 2023)	92.39	92.48	92.39	92.38
The proposed method	**96.75**	**96.75**	**97.00**	**96.80**

Table 4
The performance of the proposed classifier for each tumor types.

Tumor Type	Accuracy	Precision	Recall	F1-Score
GBM	0.98	0.98	0.97	0.974
Astrocytomas	0.97	0.97	0.97	0.97
Oligoastrocytomas	0.96	0.96	0.98	0.969
Oligodendrogliomas	0.96	0.96	0.96	0.960
Average	**0.9675**	**0.9675**	**0.9700**	**0.9680**

Fig. 6. Confusion matrix with the values obtained from the proposed method (OA: Oligoastrocytomas, OD: Oligodendrogliomas, A:Astrocytomas).

6. Discussion

Deep learning-based tools are promising to meet the needs of the precision medicine era. CNNs and vision transformers are effective in classifying neuro-pathological images. However, they have their own drawbacks. For instance, CNNs ignore global information by focusing on pixel-wise information, although they are good in the extraction of local characteristic features using several convolution and pooling layers. Vision transformers are problematics in the extraction of details and local features, although they are good in the extraction of global features using global receptive fields in the early layers (Xu et al., 2023, 2023; Lan et al., 2023; Atabansi et al., 2023; Dosovitskiy et al., 2021; Li et al., 2023). The hybrid architecture designed in this work uses benefits of these two structures. Also, the feature-combining and smart-joining modules have been designed effectively and integrated into the proposed architecture appropriately so that the architecture has the ability to retain and merge significant information.

Although ReLU is a default activation function, it has these two important drawbacks: One of them is that each unit after activation using the ReLU causes bias-shifting effects and the learning process slows down whenever the value of the mean calculated during the process is far from zero (Clevert et al., 2016). Another drawback is that neuron deaths occur in the ReLU in the case of large gradient-flows into them (Douglas and Yu, 2018). Those drawbacks have been eliminated by using LeakyReLU (Hannun et al., 2013) in this work.

Unlike typical CNN architectures where convolutional layers are linked sequentially, in a DenseNet architecture each convolutional layer is linked to all other layers which means that an input of a convolutional layer is a combination of the feature maps coming from all former layers. This encourages the spread and reuse of the features and, in this way, a reduced number of parameters are used.

The proposed architecture is efficient in multi-class classification of gliomas because: (I) The dense connections, various convolution and pooling layers in the DenseNet121 provide efficient detection and extraction of the local details. (II) The LeakyReLU provides activations without causing any biasing effects and dying neurons. (III) Global guidance of the local features, coming from the CNN stages, is provided in the transformer path using a class-token to get class-specific information. (IV) Significant global information is obtained using the MSA mechanism and MLP blocks in the transformer structures. (V) The feature-combining module reduces the misalignments between the global and local features coming from the transformer and CNN parts, respectively. (VI) The smart-joining module joins the CNNs' feature maps with the transformer's patch embeddings smartly by capturing and choosing the most valuable data to use in the prediction stage. Therefore, the proposed network achieves multi-class classification of brain tumors, namely GBM, oligodendroglioma, astrocytoma, and oligoastrocytoma with a high accuracy.

The proposed method has not been applied yet for classifications of other tissues from different medical images which can be identified as a limitation of this work.

7. Conclusions

In this work, a novel network architecture has been designed using CNN and vision transformer structures. Also, feature-combining and smart-joining modules have been designed and integrated into the architecture. The proposed hybrid network has been trained and tested to see its performance in the classification of four glioma types from histopathological images. Also, the state-of-the-art methods have been trained and tested with the same datasets to perform fair comparisons with the proposed method.

Experimental results indicate that the proposed hybrid architecture has the ability to extract local details with dense connections, convolution and pooling layers in the CNN, and global information with the vision transformer structures based on MSA and MLP blocks. Besides, global guidance of the CNN features, conversions from feature maps to patches in the feature-combining modules and joining of the patch embeddings with the CNN feature maps smartly by obtaining the most valuable data to use in the prediction stage can be performed efficiently.

It has been observed that the proposed method can achieve multi-class classifications of gliomas with a high average accuracy (96.75 %) and better performance than the other methods (Table 3). Performance analyzes and quantitative results for individual classes show that GBM tumors have been categorized with higher accuracy (98.00 %) in comparison with the other glioma types. The reason is majorly because of their more homogeneous intensities, clearer boundaries, and simpler shapes in comparison with the others. Oligoastrocytomas and oligodendrogliomas have been categorized with lower accuracy (96.00 %) compared to other glioma types because of their in-homogeneous intensity values (Table 4).

The proposed method can assist pathologists in decision-making by supporting examinations. It can increase the objectivity and diagnostic accuracy. Also, it can reduce pathologists' workload and allow them to spend more time for other complex processes. Therefore, the proposed hybrid approach is promising to overcome the issues caused by the traditional method based on microscopic examinations of glass slides and to replace it.

The proposed method will be applied for classifications of other tissues from different medical images as an extension of this work.

8. Ethics approval statement

Ethical approval was not needed since data from online sources were used in study.

Funding statement

This study has not been financially supported.

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

Data will be made available on request.

References

Aladhadh, S., Almatroodi, S. A., Habib, S., et al. (2023). An efficient lightweight hybrid model with attention mechanism for enhancer sequence recognition. *Biomolecules, 13*, 70.

Atabansi, C. C., Nie, J., Liu, H., et al. (2023). A survey of transformer applications for histopathological image analysis: New developments and future directions. *BioMedical Eng. OnLine, 22*, 1–38.

Chitnis SR, Liu S, Dash T, Verlekar TT, et al: Domain-specific pretraining improves confidence in whole slide image classification. arXiv:2302.09833 1:1-4, 2023.

Clevert, D., Unterthiner, T., & Hochreiter, S. (2016). *Fast and accurate deep network learning by exponential linear units* (pp. 1–6). Caribe Hilton, Puerto Rico: Conf. on Learning Representations.

Cluceru, J., Interian, Y., Phillips, J. J., et al. (2022). Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro-oncology, 24*, 639–652.

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing System, 34*, 3965–3977.

Dosovitskiy A, Beyer L, Kolesnikov A, et al: An image is worth 16x16 words: transformers for image recognition at scale. Int. Conf. on Learning Rep. (ICLR), virtual event, pp.1-22, 2021.

Douglas, S. C., & Yu, J. (2018). Why relu units sometimes die: analysis of single-unit error backpropagation in neural networks. In *Conf. on Signals, Syst., and Comp. California, USA* (pp. 864–868).

Gilanie, G., Bajwa, U. I., Waraich, M. M., Anwar, M. W., & Ullah, H. (2023). An automated and risk free WHO grading of glioma from MRI images using CNN. *Multimedia Tools and Applications, 82*(2), 2857.

Hafeez, H. A., Elmagzoub, M. A., Abdullah, N. A., et al. (2023). A cnn-model to classify low-grade and high-grade glioma from mri images. *IEEE Access, 11*, 46283–46296.

Hannun, A., Maas, A., & Ng, A. (2013). Rectifier nonlinearities improve neural network acoustic model. In *Workshop on Deep Learning for Audio, Speech and Language Proc., Atlanta, USA* (pp. 1–6).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks* (pp. 1–9). Honolulu, USA: IEEE Conf. on Comp. Vis. and Pattern Recognition.

Jin, L., Shi, F., Chun, Q., et al. (2021). Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers. *Neuro-oncology, 23*, 44–52.

Jose, L., Liu, S., Russo, C., Cong, C., et al. (2023). Artificial intelligence–assisted classification of gliomas using whole slide images. *Archives of Pathology & Laboratory Medicine, 147*, 916–924.

Kalaroopan, D., & Lasocki, A. (2023). MRI-based deep learning techniques for the prediction of isocitrate dehydrogenase and 1p/19q status in grade 2–4 adult gliomas. *Medical Imaging and Radiation Oncology, 67*, 492–498.

Khorasani, A., & Tavakoli, M. B. (2023). Multiparametric study for glioma grading with FLAIR, ADC map, eADC map, T1 map, and SWI images. *Magnetic Resonance Imaging, 96*, 93–101.

Komori T: The 2021 WHO classification of tumors, 5th edition, central nervous system tumors: the 10 basic principles. Brain Tumor Pathology 39:47-50, 2022.

Lan, Y. L., Zou, S., Qin, B., & Zhu, X. (2023). Potential roles of transformers in brain tumor diagnosis and treatment. *Brain-X, 1*, 1–15.

Li, Z., Cong, Y., Chen, X., et al. (2023). Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *IScience, 26*, 1–29.

Liu, Y., Liu, X., Zhang, H., Liu, J., Shan, C., Guo, Y., et al. (2023). Artificial intelligence in digital pathology image analysis. *Frontiers in Bioinformatics, 3*, 1–2.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences, 13*, 1–17.

National Cancer Institute. Genomic Data Commons Web site. Available at https://portal.gdc.cancer.gov. Accessed 28 September 2023.

Ostrom, Q. T., Cioffi, G., Gittleman, H., et al. (2019). CBTRUS Statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016. *Neuro Oncol., 21*, v1–v100.

Pei, L., Jones, K. A., Shboul, Z. A., Chen, J. Y., & Iftekharuddin, K. M. (2021). Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading. *Frontiers in Oncology, 11*, Article 668694.

Peng Z, et al: Conformer: local features coupling global representations for visual recognition. IEEE/CVF Int. Conf. Computer Vision (ICCV), virtual event, pp. 367–376, 2021.

Perry A, Wesseling P: Chapter 5 - Histologic classification of gliomas. Handbook of Clinical Neurology 134:71–95, 2016.

Pitarch, C., Ribas, V., & Vellido, A. (2023). AI-based glioma grading for a trustworthy diagnosis: An analytical pipeline for improved reliability. *Cancers, 15*, 1–28.

Prathaban, K., Wu, B., Tan, C. L., & Huang, Z. (2023). Detecting tumor infiltration in diffuse gliomas with deep learning. *Advanced Intelligent Systems, 1*, 2300397.

Shafi, S., & Parwani, A. V. (2023). Artificial intelligence in diagnostic pathology. *Diagn Pathol, 18*, 1–12.

Sharma, I., Kaur, M., Mishra, A. K., et al. (2015). Histopathological diagnosis of leprosy type 1 reaction with emphasis on interobserver variation. *Indian J Lepr., 87*, 101–107.

Sun, W., Song, C., Tang, C., et al. (2023). Performance of deep learning algorithms to distinguish high-grade glioma from low-grade glioma: A systematic review and meta-analysis. *Iscience, 1*, 1–26.

Sung, H., Ferlay, J., Siegel, R. L., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin., 71*, 209–249.

Van den Bent, M. (2010). Interobserver variation of the histopathological diagnosis in clinical trials on glioma: A clinician's perspective. *Acta Neuroapthologica, 120*, 297–304.

Wang X, Price S, Li C: Multi-task learning of histology and molecular markers for classifying diffuse glioma. arXiv:2303.14845, 1-13, 2023.

Wen, Z., Wang, S., Yang, D. M., Xie, Y., Chen, M., Bishop, J., et al. (2023). Deep learning in digital pathology for personalized treatment plans of cancer patients. *Seminars in Diag. Pathology, 40*, 109–119.

Wu, B., & Moeckel, G. (2023). Application of digital pathology and machine learning in the liver, kidney and lung diseases. *Pathology Informatics, 14*, 1–9.

Wu, P., Wang, Z., Zheng, B., Li, H., Alsaadi, F. E., & Zeng, N. (2023). AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. *Computers in Biology and Medicine, 152*, 1–10.

Xu, H., Xu, Q., et al. (2023). Vision transformers for computational histopathology. *IEEE Reviews in Biomedical Engineering, 1*, 1–17.

Younis, A., Qiang, L., Khalid, M., Clemence, B., & Adamu, M. J. (2023). Deep learning techniques for the classification of brain tumor: A comprehensive survey. *IEEE Access, 1*, 1–15.

Zhang, Y., Xiao, Y., Li, G. C., et al. (2022). How long non-coding RNAs as epigenetic mediator and predictor of glioma progression, invasiveness, and prognosis. *Seminar on Cancer Biology, 83*, 536–542.

Zhang, S., Yin, L., Ma, L., & Sun, H. (2023). Artificial intelligence applications in glioma with 1p/19q co-deletion: A systematic review. *Magnetic Resonance Imaging, 58*, 1338–1352.