Journal of Information Systems Engineering and Management
2025, 10(17s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Novel Self-Attention Framework for Robust Brain Tumor Classification

Jupalli Pushpakumari[1], Para Rajesh[2], N. Srinivas[3], Surabattina Sunanda[4], Uppula Nagaiah[5]

[1]Assistant Professor, Department of CSE-AIML & IoT VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad – 500090 Telangana, India.
pushpakumari_j@vnrvjiet.in
[2]Assistant Professor, Department of Computer Science and Engineering VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad – 500090 Telangana, India.
rajesh_p@vnrvjiet.in
[3]Assistant Professor, Department of Electrical and Electronics Engg. Vardhaman College of Engineering Hyderabad 501218 Telangana, India.
n.srinivas@vardhaman.org
[4]Assistant Professor, Department of CSE- AIDS St. Martin Engineering College Hyderabad 500014 Telangana, India.
sunandaram20@gmail.com
[5]Assistant Professor, Department of CSE- AI&ML Malla Reddy University Hyderabad 500014 Telangana, India.
nagaiah1212@gmail.com

ARTICLE INFO

Received: 08 Dec 2024

Revised: 30 Jan 2025

Accepted: 07 Feb 2025

ABSTRACT

Brain tumors rank most of the maximum extreme fitness conditions, with considerably low survival fees in superior stages, emphasizing the want for powerful remedy techniques to decorate affected person outcomes. Tumors in diverse organs, inclusive of the mind, are normally assessed the use of imaging modalities inclusive of computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound. This studies specializes in MRI scans, which can be broadly appeared because the maximum dependable approach for mind tumor detection because of their high-decision and precise imaging capabilities. However, the great facts generated via way of means of MRI scans provides giant demanding situations for guide tumor category, as it's miles time-eating and impractical for complete analysis. Additionally, best a small subset of snap shots can offer correct quantitative insights, highlighting the want for stylish computational solutions. To triumph over those obstacles, this look at proposes an automatic mind tumor category device using convolutional neural networks (CNNs) augmented with self-interest mechanisms. CNNs are specifically adept at shooting spatial functions from MRI snap shots, even as self-interest mechanisms decorate the model`s cappotential to become aware of contextual relationships throughout diverse picture regions. The proposed device classifies mind MRI scans into classes inclusive of meningioma, pituitary tumor, glioma, and non-tumor with advanced accuracy. This technique ambitions to aid clinicians via way of means of offering dependable gear for diagnosis, remedy planning, and affected person care.

Keywords: Magnetic resonance imaging, Brain tumor, Classification, Convolutional neural network, Self-attention, Meningioma, Pituitary tumor, Glioma.

1. INTRODUCTION

A mind tumor refers to an strange and frequently out of control proliferation of cells which could get up inside the mind tissue or close by regions, along with the pituitary gland, pineal gland, or shielding membranes across the mind. These tumors can disrupt everyday mind feature and effect adjoining structures, making their detection and category crucial for correct prognosis and powerful treatment. Magnetic Resonance Imaging (MRI) has emerged because the maximum dependable imaging method for figuring out and tracking mind tumors because of its cappotential to supply excessive-decision and certain pictures. Unlike computed tomography (CT) or ultrasound, MRI affords extra complete insights into mind anatomy, supporting physicians in figuring out abnormalities that could sign the presence of tumors. With its functionality to generate difficult cross-sectional pictures, MRI performs an integral function in diagnosing and monitoring the development of mind tumours.

However, studying MRI datasets gives substantial demanding situations because of the widespread quantity of facts generated. Interpreting those pictures calls for the know-how of professional radiologists, however even skilled experts can come upon errors. This underscores the want for computerized diagnostic structures able to successfully dealing with huge datasets even as keeping excessive accuracy. Manual evaluation of MRI facts is each labor-in depth and vulnerable to errors, specifically while coping with significant scientific imaging datasets. As a result, there's developing hobby in growing computer-aided prognosis (CAD) structures to assist healthcare experts in making quicker, extra correct, and dependable choices primarily based totally on MRI scans. These structures streamline the diagnostic system and decrease human error, that's specifically critical for complicated scientific situations like mind tumors.

At the center of CAD structures lies powerful picture category, which includes spotting styles in MRI scans which can imply tumor presence. The fulfillment of category relies upon on extracting significant functions from pictures, with the high-satisfactory of those functions at once influencing the model`s accuracy. In scientific imaging, conventional characteristic extraction strategies consist of depth histograms, filter-primarily based totally approaches, scale-invariant characteristic transform (SIFT), and neighborhood binary styles (LBP). These strategies purpose to seize crucial information along with texture, depth, and form which could imply tumor presence. However, those traditional strategies frequently rely upon hand made functions, which might also additionally fail to absolutely leverage the significant and complicated records found in imaging facts.

Deep mastering fashions, specifically Convolutional Neural Networks (CNNs), have revolutionized picture category via way of means of automating characteristic extraction at once from uncooked facts, putting off the want for guide engineering. CNNs are notably powerful at figuring out each neighborhood and worldwide styles in pictures [3,4], allowing them to hit upon functions that conventional strategies may overlook. These improvements have appreciably more suitable the accuracy and performance of computerized diagnostic structures, setting up CNNs as crucial equipment in present day healthcare.

A key innovation in deep mastering for picture category is the advent of interest mechanisms, which mimic human cognitive approaches via way of means of specializing in applicable elements of an picture even as ignoring inappropriate information. In neural networks, interest mechanisms permit fashions to prioritize critical functions, optimizing each reminiscence utilization and computational performance.Similar to human visual processing, these mechanisms assign different levels of importance to different regions of an image depending on their relevance.

Self-attention mechanisms are gaining importance in deep learning, especially in image classification tasks. They allow models to focus on interactions between different regions of an image, improving the network's ability to capture long-range dependencies and contextual relationships. This approach is particularly useful in applications such as brain tumor detection, where tumors may spread across multiple regions or have non-local features. Studies have shown that incorporating self-attention mechanisms improves the robustness and generality of models, allowing for more accurate and reliable tumor classification.

In summary, deep learning techniques, especially the integration of CNNs and self-attention mechanisms, have significantly advanced the field of brain tumor classification. These approaches have the potential to improve the accuracy and efficiency of automated diagnostic systems and provide medical professionals with advanced decision-making tools. These techniques enable earlier and more accurate detection of brain tumors by identifying complex patterns in MRI scans and improving model generalization, ultimately improving patient outcomes. This work highlights the potential of pairwise self-attention networks as a superior alternative to traditional convolutional networks and demonstrates its effectiveness in advancing brain tumor detection and classification [5].

2. LITERATURE SURVEY

Computer-aided detection (CAD) and diagnostic structures, which integrate synthetic intelligence (AI) and pc vision, guide radiologists in reading scientific pix to diagnose sicknesses in diverse frame regions. These structures were carried out in duties which includes the segmentation and detection of colorectal cancer [10, 11] and the class of lung cancer [12-14]. Inspired through the neural networks withinside the human mind [16], deep studying fashions combine the techniques of characteristic extraction and choice without delay into their education phases

[15]. These fashions are dependent hierarchically, wherein every layer includes a weighted mixture of capabilities from the preceding layer. The pinnacle layer represents the enter records, at the same time as the output or answer is generated withinside the backside layer. This shape allows deep studying algorithms to version complicated functions, addressing difficult troubles with minimum guide attempt as compared to standard device studying approaches. In the area of scientific photograph analysis [18], deep studying has verified superb performance, surpassing conventional device studying fashions [17].

Convolutional neural networks (CNNs) have won massive interest because of their cappotential to mechanically locate difficult capabilities and adapt to diffused versions in pix [20], presenting notably better accuracy than conventional device studying techniques [19]. A standard CNN-primarily based totally version for mind tumor class includes a couple of layers, every contributing to the extraction and class of capabilities (as proven in Fig. 1). A wellknown rule for a success generalization is to apply a dataset with as a minimum 10 instances the range of samples because the parameters withinside the version [21]. Insufficient records in the course of education can result in overfitting [22]. To conquer this, researchers [23-24] regularly make use of two-dimensional slices from third-dimensional mind MRI datasets, growing the dataset length and addressing elegance imbalance. This approach also reduces the dimensionality of the data, thus simplifying the training process.

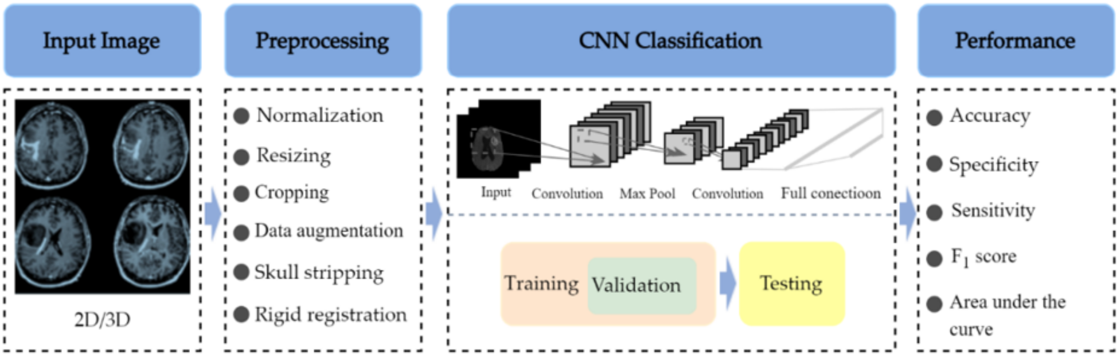


Fig. 1 The schematic view of flow of process in CNN based classification [3]

Self-attention mechanisms in computer vision have been used as a complementary component of convolutional layers so far, often integrated as an additional layer or combined with convolution to improve results. While per-channel attention models [30, 29, 28] adjust activations across different channels by applying attention weights, alternative approaches [25, 32, 27] integrate both channel and spatial attention mechanisms. To preserve the principles of convolutional feature extraction, various methods have been proposed, such as reweighting activations or modifying convolution kernel inputs [26, 29, 30, 32, 34]. Some approaches [31, 33] integrate self-attention into specific modules connected to the convolutional network.

Efforts to integrate convolutional and self-attention streams [35] showed that global self-attention applied in these combinations is not sufficient to fully replace convolution. Adaptive filter networks [36], which extend the convolutional concept, proved to be resource-intensive and unsuitable for high-resolution images and large datasets. Recent advances in [37] and [38] are closely related to our work, with the key innovation being that self-attention is applied locally across patches, rather than globally across feature maps. This localized approach reduces memory and computational overhead, enabling efficient self-attention in the first layer of high-resolution networks.

Building on these findings, our work investigates a broader range of self-attention mechanisms. Instead of using fixed scalar weights, our self-attention model computes vector attention that dynamically adapts across channels. Furthermore, we extend the notion of convolution and introduce a set of patch-wise attention operations that differ significantly from the approaches presented in [37, 38]. Our experiments highlight the scalability, efficiency, and comparable parameter and FLOP budget of the proposed self-attention model.

3. METHODOLOGY

The process of classifying medical images into predefined categories requires the ability to capture both spatial and contextual features from the image. Neural network attention mechanisms are designed to simulate human cognitive processes by focusing on relevant details and ignoring irrelevant information. This approach is particularly useful in tasks such as tumor detection and classification, where it is crucial for the model to learn how to extract important features from the most important regions of the image. Training these attention mechanisms plays a key role in the analysis of tumor images, as it is fundamental to identify the regions that are most relevant to cancer.

The preprocessing stage involves scaling, resizing, and enhancing the images. Data augmentation increases the size of the training dataset by applying various transformations to the original images. This method is particularly suitable for deep learning models, which usually require large datasets to make accurate predictions. When insufficient training data is available, augmenting the dataset can improve the model's generalization and performance on future unknown data.

3.1 Dataset

Experimental analysis was performed on a publicly available dataset from Figshare [26], a widely used resource for evaluating classification and search algorithms. The dataset contains 3064 MRI images of the brains of 233 patients diagnosed with one of three types of tumors: meningioma, glioma, or pituitary tumor. Images were acquired using a T1-CE-MRI modality that provides coronal, sagittal, and axial views. Additionally, non-tumor images were also acquired from the Br35H dataset. Table 1 provides an overview of the dataset, which includes 1426 glioma MRI images from 89 patients, 708 meningioma images from 82 patients, and 930 pituitary tumor images from 62 patients. To prepare the images for further analysis, pixel values were normalized to the range 0–1 using min-max normalization. Images were then upscaled to 224 x 224 pixels, and three color channels were generated by duplicating grayscale values to fit the input format required for the deep learning model.

Table 1. Distribution of image categories

Sno	Category	Total Samples
1	Meningioma	1042
2	Glioma	1090
3	Pituitary	1164

3.2 Overview of Self-Attention Network

The core component of a convolutional neural network (CNN) is the convolution layer, which consists of neurons that perform the convolution operation. This layer takes as input one or more 2D matrices, also called channels. The convolution operation produces several 2D output matrices, called feature maps. The number of input and output matrices depends on the number of filters used in the layer. he process required to compute a single output matrix is:

$$A_j = f(\sum_{i=1}^{N} I_i * K_{i,j} + B_j) \qquad \text{Eq.1}$$

First, the kernel matrix $K_{i,j}$ is convolved with each input matrix I_i. The resulting matrix elements are then incremented by a bias value The resulting matrix elements are augmented by bias values B_j and the sum of all the folded matrices is calculated. To generate the final output matrix A_j, a nonlinear activation function f is applied to each element of A_j. The kernel matrix acts as a local feature extractor that identifies and extracts local patterns from the input matrix. The goal of the learning process is to find a kernel matrix K that can extract discriminative features useful for tasks such as image classification. During training, the kernel matrix and biases are adjusted as connection weights using backpropagation to optimize the connection weights of the network [7]..

Pooling layers play an important role in CNNs by reducing the dimensionality of the feature map. They aggregate nearby values from the convolution output and minimize the number of neurons in the resulting feature map. Common pooling methods are max pooling and average pooling. Max pooling applies a 2x2 kernel to select the maximum value from four adjacent elements and use it to form a single value in the output matrix. In the backpropagation step, only neurons that produce the pooled output are updated, which improves the efficiency of the model and reduces overfitting.

Training neural networks can be made more effective and stable through techniques such as batch learning, momentum, and weight decay. Batch learning allows the model to process multiple samples simultaneously and adjust connection weights after processing a batch of inputs, improving both accuracy and efficiency. In our experiments, rather than adjusting the weights for each individual sample, a batch size of 128 samples was processed before updating the weights collectively. To improve the convergence process, mechanisms such as momentum[8] and weight decay[9] were introduced. The weight update can be defined as follows:

$$\Delta\omega_i(t+1) = \omega_i(t) - \eta\frac{\partial E}{\partial w_i} + \alpha\Delta\omega_i(t) - \lambda\eta\omega_i \qquad \text{Eq. 2}$$

$\omega_i(t) - \eta\frac{\partial E}{\partial w_i}$ represent the conventional back-propagation term in which $\omega_i(t)$ is the current weight vector and $\frac{\partial E}{\partial w_i}$ denotes the error gradient with respect to the weight vector. The η denotes the learning rate which usually controls the rate of convergence. The $\alpha\Delta\omega_i(t)$ represent the momentum term wherein the α denotes the momentum rate. The λ represents the load decay fee which enables to lessen the threat of version overfitting. Each studying generation consequences in a tiny discount or decay of the load vector toward zero, helping withinside the stabilization of the studying technique. In our experiment, we set the studying fee to be 0.001, the momentum fee to be 0.9, and the decay fee to be 0.01. These studying parameters have been selected the use of a grid seek mechanism.

In convolutional neural networks for picture recognition, the convolution layers serve number one functions. The first feature is function extraction, wherein the convolution operation aggregates facts via way of means of combining values from areas blanketed via way of means of the kernel. The 2d feature is function transformation, which applies nonlinear mappings observed via way of means of linear mappings. These differences partition the function area into smaller segments, growing complex, piecewise mappings. A key perception is that function aggregation and transformation may be dealt with separately. Once function aggregation is handled, perceptron layers can independently technique every function vector (similar to every pixel) for transformation. The perceptron layer plays pointwise operations, remodeling functions thru a mixture of linear mappings and nonlinear scalar functions. As a result, the layout of deep studying-primarily based totally picture type focuses usually on function aggregation. The convolution operator makes use of a fixed, pre-skilled kernel with set weights to linearly integrate function values from neighboring areas. These weights are regular and do now no longer extrade relying at the enter picture. Moreover, the range of parameters will increase linearly with the range of aggregated functions, as every place calls for a completely unique weight vector.

By integrating function aggregation thru self-interest with function transformation through element-sensible perceptrons, enormously green picture type architectures may be achieved. Self-interest expands the CNN`s receptive area via way of means of thinking about big kernel sizes with out substantially growing computational demands. Attention mechanisms have end up vital for deep studying fashions that want to seize worldwide dependencies. Specifically, self-interest (or intra-interest) computes responses for a given place in a chain via way of means of thinking about all different places in the identical sequence. The pairwise self-interest may be expressed mathematically as follows;

$$y_i = \sum_{j\in R(i)} \alpha(x_i, x_j) \odot \beta(x_j) \qquad \text{Eq. 3}$$

where $\odot$ denotes the Hadamard product, i represents the spatial index of the feature vector denoted by x_i, and $R(i)$ denotes the local print of the aggregation which is nothing but the set of all indexes that indicates which feature vectors are aggregated for constructing the output feature vector y_i. The feature vector represented by $\beta(x_j)$ are generated by the functionβ and these features are aggregated by the adaptive weight vectors denoted as $\alpha(x_i,x_j)$. The computed weights $\alpha(x_i,x_j)$ are used for combining the transformed features $\beta(x_j)$. The function α can be decomposed as follows;

$$\alpha(x_i,x_j) = \gamma(\delta(x_i,x_j)) \qquad \text{Eq. 4}$$

A single vector which represents the features x_i and x_j will be given as output by the relation function δ. The γ function later maps this single vector which is combined with the transformed features generated by$\beta(x_j)$.The dimension of the output from the γ function will not be same as the transformed features as the attention weights were shared among a set of channels.An unique feature of the pair-wise attention mechanism described above is that the feature vectors are processed independently and the computed weights does include only information related to location denoted by i and j. To incorporate the spatial information, the feature maps are augmented as follows;

Step 1. The horizontal and vertical coordinates across the feature map are normalized to a range of [-1, 1] in each dimension.

Step 2. The normalized coordinates are then passed through a trainable linear layer, which maps them to an appropriate range for each layer mmm in the network (i.e. m=network layers).

Step 3. The output of the linear mapping function produces a 2D feature p_i for every location within the feature map.

Step 4. The relative location information is encoded by calculating the difference $p_i–p_j$, which is used to augment the feature maps with spatial context .

Step 5. Augment the weight vectors $\alpha(x_i,x_j)$ by concatenating the $[p_i - p_j]$ before γ function.

The self-attention mechanism described above was used to construct a residual block which shall implement both feature aggregation and transformation. The input feature vector is sent via two processing streams. The left unit asses the weights of the attention layer through mapping function δ and a mapper denoted as γ. The right unit reduces the dimensionality of the input by applying a linear transformation β. Then the output of the right and left units are aggregated through a Hadamard product. Finally the combined features are normalized and expanded to their original dimension C.

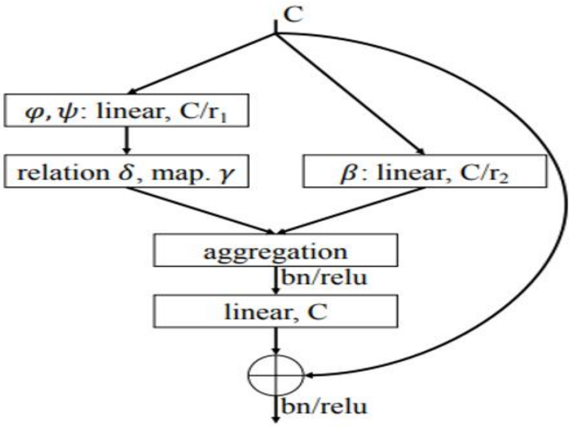


Fig. 2 Schematic view of a self-attention block [2]

3.3 Network Architecture

The proposed network architecture is inspired by residual networks and follows established methods [12]. It integrates 10 self-attention blocks with different resolutions, with ResNet50 serving as the core element of the design. At the heart of this architecture is the self-attention mechanism underlying the Self-Attention Network (SAN). The SAN consists of five stages, each operating at a different spatial resolution, resulting in a 32x reduction in resolution. Each stage contains multiple self-attention blocks, and transition layers are used to connect successive stages. These transition layers reduce the spatial resolution while increasing the channel dimensionality. The final output of the last stage passes through a classification layer consisting of global average pooling, a linear layer, and a softmax function to facilitate classification. The transition layers are designed to reduce computational complexity by expanding the receptive field and reducing the spatial resolution. These layers use batch normalization, ReLU activation function, 2x2 max pooling with step size 2, and linear transformation to increase the channel dimensionality. The architecture of the Self-Attention block is shown in Figure 2.

3.4 Performance Metrics

The performance of the classifier developed for the multi-class classification task was assessed using several key metrics:

- **Matthew's Correlation Coefficient (MCC)**: Based on a confusion matrix typically used to evaluate binary classifiers, MCC is a reliable metric for multiclass models, especially in the presence of class imbalance. MCC values range from -1 to 1, with values closer to 1 indicating good performance and values closer to 0 indicating poor classification results.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \text{Eq. 5}$$

- **Log Loss Function: Log Loss Function:** This error function, also known as cross-entropy loss, measures the discrepancy between the actual labels and the predicted probabilities. The model that minimizes the log-loss function is considered optimal. Unlike other metrics that rely solely on predicted labels to evaluate performance, log-loss takes into account the model's confidence in its predictions. This metric assesses uncertainty by calculating a loss based on probability values, penalizing cases where the model assigns a low probability to the correct class. This increases confidence in the model's predictions and improves overall accuracy.

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \qquad \text{Eq. 6}$$

where M – number of classes present in the dataset

Y – a binary value indicating the correct classification for observation 'o'

P – predicted probability score indicating observation o is of type c.

- **Cohen's Kappa Coefficient**: Cohen's Kappa Coefficient (k) is a statistical metric used to assess the consistency of agreement for categorical (qualitative) items. It applies to both inter-rater reliability (assessing agreement between different raters) and intra-rater reliability (assessing consistency within the same rater over time). Unlike simple percentage agreement, kappa provides a more reliable and precise measure because it takes into account the possibility of agreement by chance. Higher kappa values indicate better agreement, whereas lower kappa values indicate poorer agreement

$$k = \frac{p_o - p_e}{1 - p_e} \qquad \text{Eq. 7}$$

The time period po refers back to the located concordance among the real and anticipated values, representing the share of successfully categorised times withinside the confusion matrix. It is computed because the sum of the diagonal elements (real positives and real negatives) divided through the whole variety of times. On the alternative hand, pe denotes the possibility of settlement taking place basically through chance. It is the probability that the real values and expected values will fit randomly. The distinction among po and pe is frequently used to assess the general overall performance and accuracy of a classifier.

4. EXPERIMENTS AND RESULTS

The enter photo length for the community became standardized to 224x224 pixels, performed through resizing all pics the usage of bi-cubic interpolation. The version underwent education for one hundred epochs, with the dataset break up into non-overlapping subsets for education and validation. To make sure an most beneficial break up among education and trying out data, k-fold cross-validation became used. To similarly decorate the education dataset, diverse photo augmentation strategies have been applied, inclusive of flipping, rotation, and elastic deformation. The community`s weights have been initialized randomly at the start of education, with the studying charge set at 0.0001 and the batch length set to 16. While flipping and rotation preserved the unique form and length of the tumors, elastic deformation brought pixel rearrangements that would doubtlessly distort the tumors' structure. This deformation should purpose the tumors to lose their unique visible look or be misclassified as malignant because of the altered context. The effects indicated that elastic augmentation stepped forward version overall performance to 88.64%, while flipping and rotation furnished even higher overall performance with an accuracy of 91.83%. The test additionally evaluated the version's overall performance on pics that have been turned around and flipped. Using the Figshare dataset, the pics have been manipulated through rotating them 90° clockwise and appearing a horizontal flip. This served as zero-shot trying out, as those adjustments have been now no longer a part of the version's education process. The effects, proven in Table 2, aid the speculation that pairwise self-interest, a set-primarily based totally operation, might reveal higher resilience to those adjustments than convolutional networks or patch-clever self-interest fashions. The findings verify this speculation, displaying that even though all networks have been suffering from area shifts, the pairwise self-interest fashions exhibited more robustness as compared to their convolutional and patch-clever self-interest counterparts.

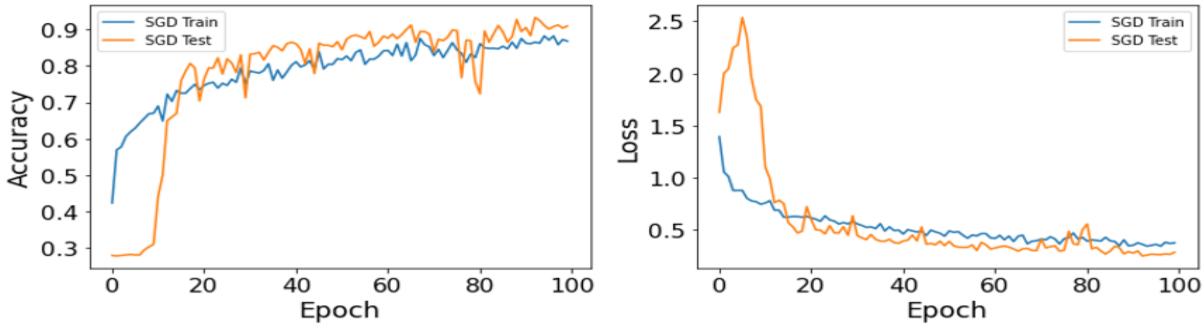
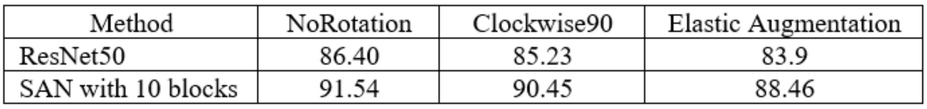


Fig. 3 Analysis of Accuracy and Loss for Attention based classifier with SGD optimization

Table 2. Performance comparison of proposed self-attention based model with Conv. ResNet model



Method	NoRotation	Clockwise90	Elastic Augmentation
ResNet50	86.40	85.23	83.9
SAN with 10 blocks	91.54	90.45	88.46

The category overall performance of the attention-primarily based totally version became evaluated the usage of the Matthews Correlation Coefficient (MCC), which confirmed advanced overall performance throughout all 4 classes of the confusion matrix: genuine positives, fake negatives, genuine negatives, and fake positives. This sturdy overall performance became steady with the distribution of each advantageous and terrible factors withinside the dataset, making MCC a greater dependable and balanced metric. Additionally, the excessive price of the kappa coefficient suggests that the proposed version continues sturdy overall performance, even if confronted with sizeable magnificence imbalance withinside the schooling dataset.

Table 3. Summary of Performance Metrics (SGD with LR = 0.0001)

Matthew's Correlation Coefficient	Cohen's Kappa coefficient	Log Loss
0.9406	0.9385	**0.1632**

5. CONCLUSION

In this study, we investigate the effectiveness of an advanced image classification model based entirely on the self-attention mechanism. Unlike traditional convolutional models, which utilize convolutions to extract spatial features, this model uses pairwise self-attention, a special operation that allows the network to focus on relationships between distant parts of an image. The self-attention mechanism allows the model to capture contextual dependencies across different domains, making it highly effective in tasks that require long-distance relationships or dependencies. The model uses vector attention to tune parameters across both spatial dimensions and channels, allowing the network to learn more complex and subtle patterns in the data. By focusing on pairwise interactions, the model is able to better understand the meaning of different image regions in relation to each other, facilitating a more comprehensive understanding of images. This vector-based approach also improves the adaptability of the model to different kinds of visual data, with the advantage of capturing complex structures that may be missed by traditional methods.

The experiments conducted in this study yielded several notable results that highlight the effectiveness of self-attention in image classification tasks. Compared to convolutional networks, models based solely on pairwise self-attention performed better in a variety of tasks. These findings call into question the traditional view that convolutional layers are essential to achieve high performance in deep learning-based computer vision tasks. Instead, the self-attention mechanism offers a promising alternative that may rival or even surpass the performance of convolutional models.

One of the main advantages of the self-attention mechanism is its ability to function independently of the spatial arrangement of the data. This capability gives self-attention networks permutation and cardinality invariance, meaning that the network can process and classify images regardless of how the elements in the network are arranged. This property gives self-attention networks a significant advantage as they can process complex and dynamic visual data in ways that convolutional networks have difficulty with, especially when dealing with variable input sizes and irregular patterns. The results of this study indicate that the self-attention mechanism is a promising alternative to traditional convolutional approaches, offering structural advantages that may result in comparable or even better performance in image classification tasks. The ability of self-attention models to capture long-range dependencies and maintain invariance to permutations and cardinality opens new possibilities for deep learning-based computer vision, especially in tasks where understanding global relationships between image regions is important. As shown in this work, self-attention mechanisms represent a promising direction for improving the capabilities of image classification models and may lead to more efficient and powerful solutions in computer vision..

REFERENCE

[1] Abiwinanda, Nyoman, et al. "Brain tumor classification using convolutional neural network." *World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1)*. Springer Singapore, 2019.

[2] Zhao, Hengshuang, JiayaJia, and VladlenKoltun. "Exploring self-attention for image recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[3] Xie, Yuting, et al. "Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): Review, challenges, and future perspectives." *Diagnostics* 12.8 (2022): 1850.

[4] Ari, Ali, and DavutHanbay. "Deep learning based brain tumor classification and detection system." *Turkish Journal of Electrical Engineering and Computer Sciences* 26.5 (2018): 2275-2286.

[5] Li, Zhenliang, et al. "Deep multi-instance learning with induced self-attention for medical image classification." *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020.

[6] Zhao, Hengshuang, JiayaJia, and VladlenKoltun. "Exploring self-attention for image recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[7] Takahashi, Ryo, Takashi Matsubara, and Kuniaki Uehara. "A novel weight-shared multi-stage CNN for scale robustness." *IEEE Transactions on Circuits and Systems for Video Technology* 29.4 (2018): 1090-1101.

[8] Kan, Tao, et al. "Convolutional neural networks based on fractional-order momentum for parameter training." *Neurocomputing* 449 (2021): 85-99.

[9] Hernández-García, Alex, and Peter König. "Do deep nets really need weight decay and dropout?." *arXiv preprint arXiv:1802.07042* (2018).

[10] Figueiredo, P.; Figueiredo, I.; Pinto, L.; Kumar, S.; Tsai, Y.; Mamonov, A. Polyp detection with computer-aided diagnosis in white light colonoscopy: Comparison of three different methods. Endosc. Int. Open 2019, 7, E209–E215.Diagnostics 2022, 12, 1850 42 of 46

[11] Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput. Biol. Med. 2021, 137, 104815.

[12] Gong, J.; Liu, J.Y.; Sun, X.W.; Zheng, B.; Nie, S.D. Computer-aided diagnosis of lung cancer: The effect of training data sets on classification accuracy of lung nodules. Phys. Med. Biol. 2018, 63, 035036.

[13] Nishio, M.; Sugiyama, O.; Yakami, M.; Ueno, S.; Kubo, T.; Kuroda, T.; Togashi, K. Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PLoS ONE 2018, 13, e0200721.

[14] Tian, Q.; Wu, Y.; Ren, X.; Razmjooy, N. A new optimized sequential method for lung tumor diagnosis based on deep learning and converged search and rescue algorithm. Biomed. Signal Process. Control 2021, 68, 102761.

[15] Schmidhuber, J. Deep learning in neural networks: An overview. Neural Netw. 2015, 61, 85–117.

[16] Hu, A.; Razmjooy, N. Brain tumor diagnosis based on metaheuristics and deep learning. Int. J. Imaging Syst. Technol. 2021, 31, 657–669.

[17] Tandel, G.S.; Balestrieri, A.; Jujaray, T.; Khanna, N.N.; Saba, L.; Suri, J.S. Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm. Comput. Biol. Med. 2020, 122, 103804.

[18] Shen, D.; Wu, G.; Suk, H.I. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 2017, 19, 221–248.

[19] Tandel, G.S.; Balestrieri, A.; Jujaray, T.; Khanna, N.N.; Saba, L.; Suri, J.S. Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm. Comput. Biol. Med. 2020, 122, 103804.

[20] Yasaka, K.; Akai, H.; Kunimatsu, A.; Kiryu, S.; Abe, O. Deep learning with convolutional neural network in radiology. Jpn. J. Radiol. 2018, 36, 257–272.

[21] Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. 2018, 19, 1236–1246.

[22] Deepak, S.; Ameer, P.M. Brain tumor classification using deep CNN features via transfer learning. Comput. Biol. Med. 2019, 111, 103345.

[23] Ge, C.; Gu, I.Y.H.; Jakola, A.S.; Yang, J. Deep semi-supervised learning for brain tumor classification. BMC Med. Imaging 2020, 20, 87.

[24] Gab Allah, A.M.; Sarhan, A.M.; Elshennawy, N.M. Classification of Brain MRI Tumor Images Based on Deep Learning PGGAN Augmentation. Diagnostics 2021, 11, 2343.

[25] Long Chen, Hanwang Zhang, Jun Xiao, LiqiangNie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017. 2

[26] Jifeng Dai, Haozhi Qi, YuwenXiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, 2017.

[27] Jun Fu, Jing Liu, HaijieTian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, 2019.

[28] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In NeurIPS, 2018

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, 2018.

[30] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In CVPR, 2017.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, 2018.

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In ECCV, 2018.

[33] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and JiayaJia. PSANet: Point-wise spatial attention network for scene parsing. In ECCV, 2018.

[34] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets v2: More deformable, better results. In CVPR, 2019.