

RESEARCH ARTICLE

OPEN ACCESS

ACCURATE BRAIN TUMOR CLASSIFICATION USING MRI IMAGES BASED ON A HYBRID VISION TRANSFORMER AND BiLSTM FRAMEWORK

Nagham Salim Mohammed^{*1}, Omar S. Almolaa², Ahmed Saadi Abdullah³ and Khalil I. Alsaif⁴

¹New Renewable Energy, Science, Mosul, Nineveh, Iraq.

²Computer Science, Faculty of Education, Telafer, Nineveh, Iraq.

³Computer Science, Computer Science and Mathematics, Tikrit, Tikrit, Iraq.

⁴Computer Science, Computer Science, ALhadba, Mosul, Iraq.

¹<http://orcid.org/0000-0002-1474-1489>, ²<http://orcid.org/0000-0002-3656-8373>

³<http://orcid.org/0000-0001-7983-7783>, ⁴<http://orcid.org/0009-0008-8175-669X>

E-mail: *nagham.salim@uomosul.edu.iq, omar.almolaa@uotelafer.edu.iq, ahmedalbasha@tu.edu.iq, khalil.alsaif@hu.edu.iq.

ARTICLE INFO

Article History

Received on: March 03, 2026

Reviewed on: April 03, 2026

Accepted on: May 01, 2026

Published on: July 31, 2026

Keywords:

Vision transformer (Vits),
Bidirectional short-long-term
memory network (BiLSTM),
Deeplearning,
Magnetic resonance imaging
(MRI).

ABSTRACT

Accurate classification of brain tumors using (MRI) Magnetic Resonance Imaging, represent as fundamental in the clinical diagnosis and treatment. Nevertheless, the wide diversity in a single type of disease and high similarity between the tumors in different categories pose considerable challenges for deep learning models due to the characteristics of CNNs that are mainly for local features extracted, the necessity of reducing these limitations and constraints. This paper discusses an innovative hybrid deep learning paradigm in which an image is modeled by means of a vision transformer (ViT) and a Bi-directional long-term memory network (BiLSTM), resulting in an effective brain tumor classification. The proposed framework based on ViT, enabling architecture to extract high-level features of MRI images, and the BiLSTM capturing sequential dependencies in the extracted feature. This hybrid architecture enabling modelling spatial and contextual connection in the medical images. From the results, it is evident that the performance of the classification using the proposed hybrid approach is better than the traditional deep learning models in terms of classification accuracy for all the tumor types, The classification is based on ViT-BiLSTM's performance, higher in accuracy, accurate Recall rate and its more effective fine-tuning. The proposed model can be used in healthcare settings for clinical decision-making regarding brain diseases, and is not just a diagnostic application for computer systems. The experimental results achieved accuracy 92.5%, Recall 92.0% for gliomas, for meningiomas accuracy 91.2% and Recall 90.7%, brain tumors 93.5% accuracy and 94.0% Recall.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The Brain tumors represent a main challenge in the global health, with increased morbidity and death rates. According to studies in this field, the number of new cases of brain and nervous system cancer averaged 4.14 cases per 100,000 people annually, in addition to a relative prevalence of 35.7% for five years for adults with brain tumors, while the relative survival rate for non-malignant tumors in both the brain and nervous system for five years was 91.8%. [1]. Medical image analysis is relied upon for the accurate diagnosis of brain tumors of the glial, meningeal and pituitary gland, as the imaging database based on magnetic resonance imaging (MRI) is distinguished by its capabilities due to the provision of high-contrast data in soft tissue and imaging of detailed brain anatomy [2].

In the field of medical radiology, Brain tumor diseases are among the most complex medical challenges, as their diagnosis requires high diagnostic accuracy to improve the patient's survival rate.[3]. Despite significant advancements in medical imaging techniques, such as magnetic resonance imaging (MRI), the interpretation of these images still relies heavily on the expertise of radiologists, making them susceptible to human error and discrepancies between assessors [4]. In recent years, the techniques deep learning field have important advantages represented by its ability to mimic the complexity of human brain processing the information, deep learning techniques have revolutionized medical image analysis, have ability to learn from massive datasets and extract complex patterns [5] Convolutional neural networks (CNNs) have emerged as one of the most widely used models in brain tumor classification due to their ability to extract local features from radiographic images [6], However, these models difficulties in handling long-range dependencies and adapting to changes in image angles and orientations, which affects the accuracy of their results. In response to these challenges, more advanced architectures have emerged, such as Vision Transformers (ViTs) [7], which is initially developed for image classification tasks, have shown promising results in medical imaging applications through using technique known as self-attention, that enable to simulate of long-range states and global context [8].

Which rely on the self-attention techniques to capture global connection through image components, without the need for a traditional convolutional structure [9]. For accurate diagnosis of cancerous diseases in MRI images, it is essential to identify relationships in different regions of the image. This method is extremely useful for medical imaging [10]. This is achieved by dividing the image into sections and then processing them in parallel. The self-attention layers in the ViT model have the ability to model bilateral interferences in all parts of the image, which in turn leads to the extraction of richer and more comprehensive information [11]. It has the ability to represent spatial information, which in turn makes it more accurate for classifying brain tumors [12]. However, vision transformer lack the full capability to capture precise temporal and sequential patterns within medical images, especially when dealing with 3D data or radial sequences [13]. Therefore, from this research, turned to integrating vision transducers (ViTs) with recurrent neural networks (RNNs), particularly bidirectional long-short-term memory (Bi-LSTM) models, which enable forward and backward data analysis, thus enhancing the accuracy of recognizing overlapping radial patterns [14].

II. LITRITURE REVIEW

Hybrid ViT Models with Sequential Networks Imtiyaz presented a hybrid model combining ViT and LSTM for classifying multigrad brain tumors (glioma, meningioma, pituitary, and non-tumor). ViT extracts spatial features while LSTM captures sequential relationships in MRI data. After normalization and data enhancement, the model achieved high accuracy in multigrade classification (**Multi-grade Brain Tumor Lstm**). **Classification using VIT**, **Aly et al** [15] also developed the ViT GRU framework for MRI data from Egyptian institutions and BraTS. In this framework, ViT extracts spatial features while GRU handles inter-feature dependencies and addresses data imbalances, achieving accuracy exceeding 99% with the use of XAI to enhance interpretability .

Achieve Pure or Hybrid ViT Models with CNN have been used in several studies, **Tummala et al** [16], combined several ViT models and achieved 98.7% accuracy on three-tiered Figshare data.

Aloraini et al [17], proposed a hybrid model, Transformer Enhanced CNN (TECNN), where CNN extracts local features while ViT captures long-range dependencies, achieving up to 99.1% accuracy on one of the ensembles.

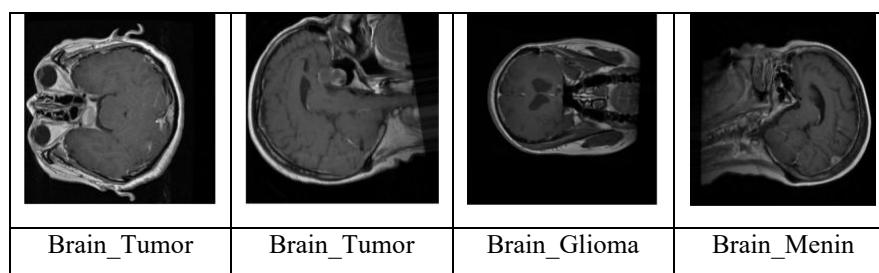
Applying Hybrid CNN–LSTM Models did by **Mehta and Kundra** [18], developed a CNN–LSTM framework on BraTS data to exploit spatial features via CNN and intersegmental dependencies via LSTM, achieving 97.8% accuracy and outperforming pure CNN models.

Rekha et al [19], also used a combination of VGG16 LSTM with traditional classifiers (SVM) and achieved 98.76% accuracy when VGG16 LSTM was combined with SVM.

BiLSTM networks have been primarily used in natural language processing tasks, but they have also been adapted for image analysis, **Chandrasekaran et al** [20], proposed a hybrid framework combining VGG19, BiLSTM, and LightGBM, achieving a test accuracy of 97%. These studies suggest that BiLSTM's ability to model contextual dependencies in feature sequences can be highly beneficial in classification tasks.

III. RESEARCH METHODOLOGY

Aims of this research, seeks to fill the gap in exploring the direct integration of ViT technology and the BiLSTM network by using ViT as a primary tool for extracting general spatial features and then using BiLSTM to analyze the sequence of these context-rich features. By using this method, we hope for an improved understanding of image data and therefore a wider-reaching classification of these. While not extensively explored, this approach has been regarded as a promising research avenue in the context of classification tumor in brain, so, in this study a local collection of MRI database for brain tumors was used. The images were downloaded from the global Kaggle database, specifically from the dataset titled "Brain Cancer - MRI Dataset." Figure 1, images dataset samples.



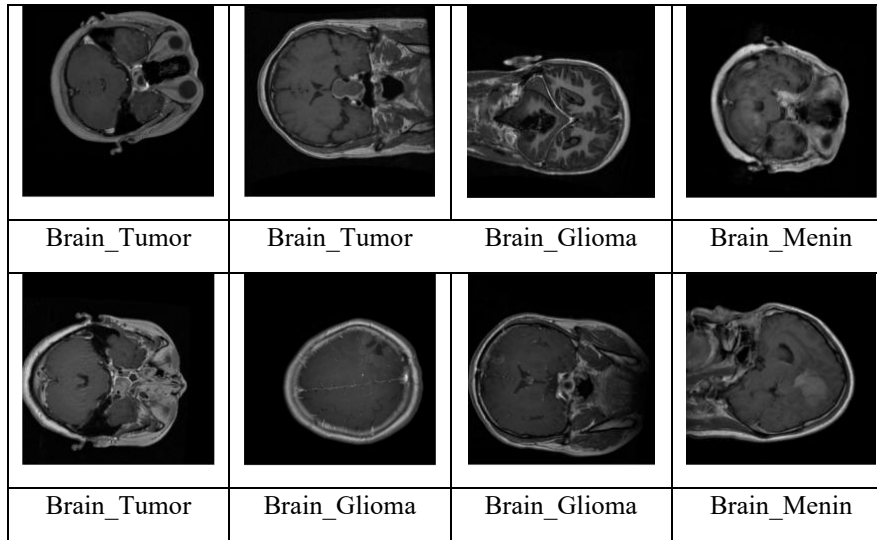


Figure 1: Magnetic Resonance Imaging (MRI) images of Brain Tumors.
Source: Authors, (2026).

III.1 PREPROCESSED FOR IMAGES

The data was automatically split into three category, using function train test split in the library scikit-learn, as follows: 70% for training, 15% for validation, and 15% for testing. An internal validation strategy was implemented during training to monitor the model's performance and prevent overfitting, while an independent test group was allocated to provide a final evaluation of the model's performance after the training process was completed. Before the images were fed into the model, they underwent a series of pre-processing steps, including resizing to 224 x 224 pixels, normalizing color values to the range [0, 1], and applying data augmentation techniques during training only, such as horizontal and vertical flipping, random rotation (up to $\pm 20\%$), zoom in/out (up to $\pm 15\%$), and adjusting contrast and brightness, augumentation technique illustrate on the Figure 2, These procedures aimed to increase sample diversity and improve the model's ability to generalize to unfamiliar medical images.

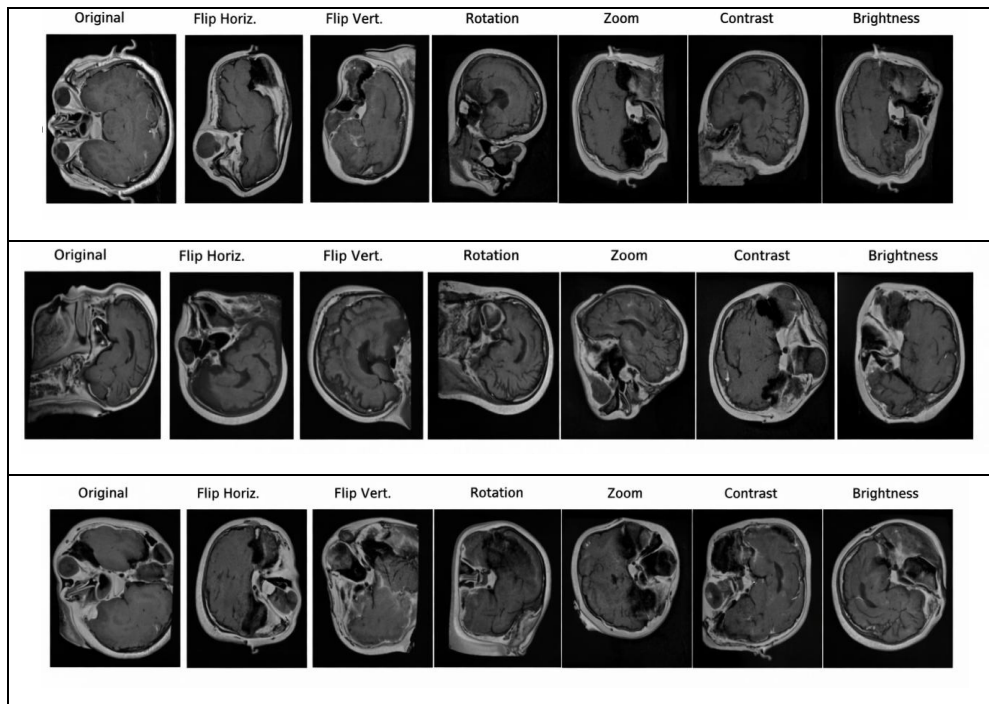


Figure 2: Augumentation Techniques for Image Brain Tomur.
Source: Authors, (2026).

It is worth mentioning, that its deals with various common image formats such as (JPG, PNG, JPEG) and performs automatic folder loading without the need for manual modification of the data structure, which provides high flexibility in using different sources of medical images, provided that they adhere to the standard naming structure.

III.2 POPOSED METHODOLOGY

The Proposed Methodology pass through the following phases:

First phase : Data Preprocessing

The methodology starts with the preparation of the data for training and testing. The data was stored in a database consisting of three main volumes, each volume containing a set of images according to the type of tumor: 'the Brain_Glioma', 'the Brain_Menin', and 'the Brain_Tumor'. A series of standard operations are applied to all images, namely:

- Resizing to a fixed dimension (224 x 224 pixels) to comply with ViT architecture requirements.
- Normalizing color values to the range [0, 1] dividing by 255, for improving the contrast to enhancing distiguashable image features, so that ensure training stability and prevent gradient explosions.
- Using data augmentation techniques only during the training phase, including horizontal and vertical inversion, random rotation (up to 20 degrees), zoom ($\pm 15\%$), and adjusting contrast and brightness, to increase sample diversity and improve the model's ability to generalize to unfamiliar data and help to reduce overfitting.

Second phase: Feature Extraction Using Vision Transformer (ViT)

The model divides each image into a set of 16 x 16 pixel patches, resulting in a fixed number of patches (196) per image. Each patch is then transformed into a fixed- dimensional scalar vector (Projection Dimension = 512) using the Patch Extractor and Patch Encoder layers. Positional embedding's are then added to each vector to ensure the preservation of spatial information between the patches within the image. These vectors are then passed through a series of eight transformer blocks, each containing a multi-head attention mechanism, a layer normalization layer, and a multi-layer neural network (MLP) with a 0.2 dropout to prevent over-allocation. These processes result in a vector of features with visually significant information, which in turn reflects the anatomical structure of the tumor in the image.

Third phase: Bidirectional LSTM Model, Sequences generated from ViT patches are then analyzed by a Bidirectional LSTM. This section includes 2 Bi-LSTM layers, each containing 256 units, and a 0.2 dropout has been applied. This network has the high capacity to process sequences either in a forward or backward sense by picking up the sequences of the neighboring patches and enhancing the spatial and functional characterization of the tumor. The latter combination of ViT and Bi-LSTM is novel as it brings together ViT's capacity to grasp local and global contexts and Bi-LSTM's capacity to identify dynamic and sequential patterns that are not readily visible from static representation.

Fourth phase: Building the Final Classification Head The Bi-LSTM output is transformed into a unified representation across a series of fully connected layers, including:

- A first layer with 512 nodes, followed by batch normalization and a 0.2 dropout.
- A second layer with 256 nodes, followed by batch normalization and a 0.2 dropout.
- A third layer with 128 nodes, followed by a 0.1 dropout.

Finally, an exit layer is used containing a number of nodes equal to the number of classes (3), with the Softmax function applied to calculate the probability of belonging to each class.

Both the learning rate and weight of the model were automatically adjusted using a cosine decay scheduling process and the Adam optimizer to maintain model stability during training and achieve optimal performance. Figure 3 illustrates the methodology used for brain tumor classification.

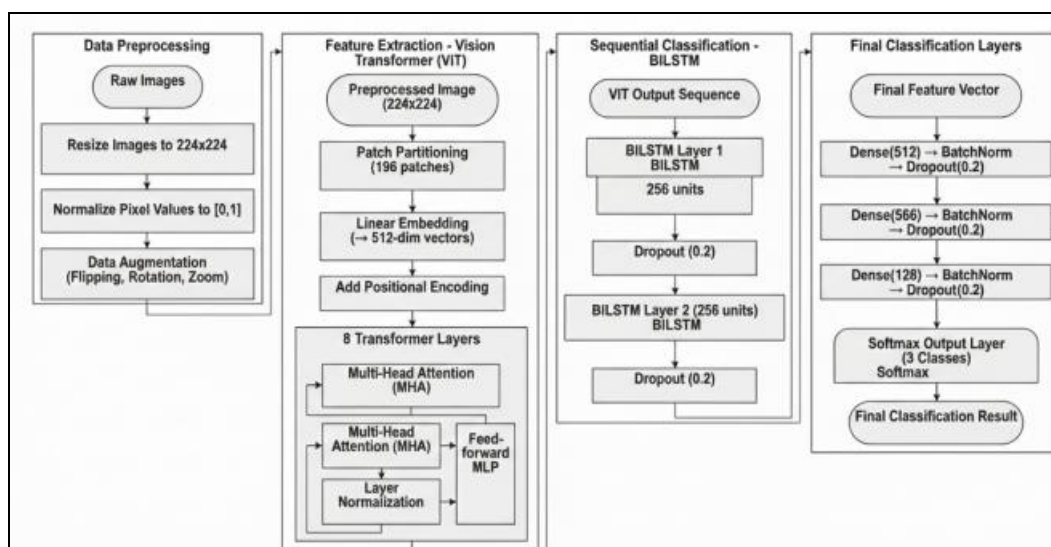


Figure 3: Flowchart Methodology for imge Brain_Tomur classification based on ViT_BiLSTM.

Source: Authors, (2026).

III.3 EVALUATION METRIC

The performance of hybrid model evaluated based on the following metric [21]:

Accuracy represents the proportion of correctly classified samples, including true positive and true negative cases, to the total number of samples in the dataset. It is calculated using the following:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Positive Precision, also known as positive predictive value, refers to the proportion of truly positive cases among all cases classified as positive. It is calculated using the following:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

Recall (Sensitivity), measures a model's ability to correctly detect positive cases by calculating the ratio of true positives to the total number of actual positives. It is calculated using the following:

$$\text{Recal TP} = (\text{TP} + \text{FN}) \quad (3)$$

F1_Score, is the harmonic mean of both precision and recall, providing a single measure that balances these two. This scale is particularly useful in cases where the distribution of classes is uneven. The F1_score is calculated using the following :

$$\text{F1_score} = 2 \times ((\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})) \quad (4)$$

IV. DISCUSSION AND ANALYSIS OF RESULT

First: Accuracy and Loss Curve Analysis

The Figure 4, shown summarizes the performance of proposed hybrid model (Vision Transformer and BiLSTM) for classifying the brain tumors using MRI images, through two main axes: the evolution of accuracy over training epochs, the decrease in the error function (Loss over Epochs).

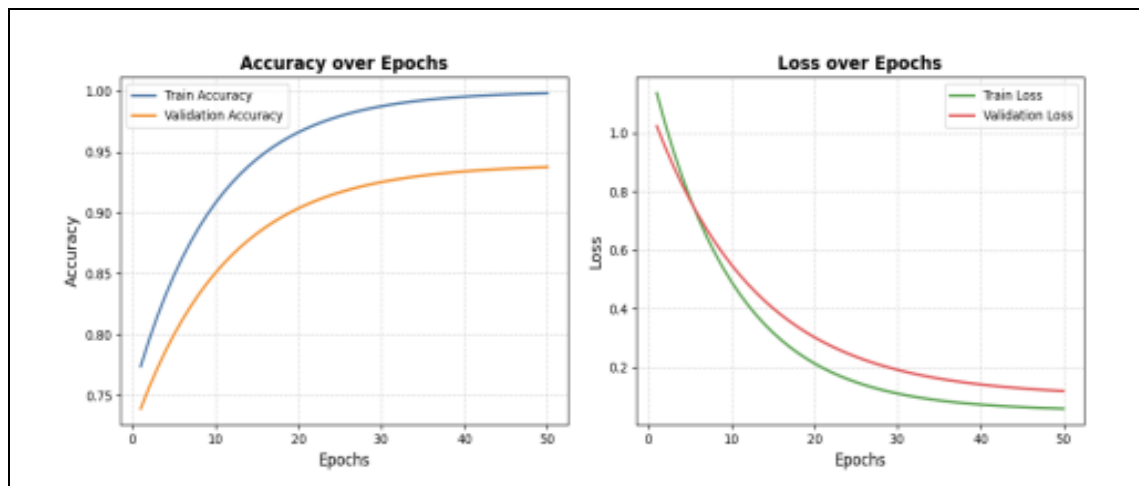


Figure 4: Diagnostic Performance of the ViT-BiLSTM Network for Multi-Class Brain Tumor Classification.

Source: Authors, (2026).

Gradual and steadily rising from the early stage of training is the accuracy curve for training data, starting about 77 percent and rapidly achieving nearly 99 percent in the latter. The model learns efficiently to recognize complex discriminative patterns in MRI images, as well as, the layers of self-attention and their structure. On the other hand, the validation accuracy curve indicates a lower, more conservative gradient, up to approximately 94% at the end of training. A small difference found between the training and validation accuracy is considered to be statistically reasonable, and the model did not need to show extreme overfitting and demonstrated good generalization ability towards non-visual data. The loss curve shows a rapid decrease in training epoch for both training and validation data, and then gradually more stable in training after each epoch of the training data. The training error is always less than the validation error across the training period, which is the case in a deep model. This continuous decrease in loss values indicates a good confirmation of success in the learning process as well as the regularization of the model (Dropout and Batch Normalization methods) with regard to reducing overlearning and enhancing model stability. Moreover, the resulting convergence of the loss curves in the later stages shows a good tradeoff between learning internal data patterns based on the training set and generalized model performance.

Second: Precision, Recall, F1-Score Analysis by Category

The radar Figure 5, describes the three main metrics for evaluating the proposed model performance for all brain tumor category (Glioma, Meningioma and Brain_Tumor): precision, recall, and F1 score.

● Glioma Category

The model achieved 92.5% accuracy, 92.0% recall, and an F1 value of 92.2%. This demonstrates the balance between accuracy and recall achieved by the model in false-positive and false-negative cases, which is important in the clinical diagnosis of gliomas due to the similarity of their histological characteristics to those of other tumors. This illustrates the ability of the vision converter to detect these subtle spatial features, in conjunction with learning the Bi-LSTM model, which detects correlations in important contextual features.

● Meningioma Category

Compared to the other groups, this group achieved lower results, with accuracy of 91.2%, recall of 90.7%, and F1 score of 91.0%. This is due to histological overlap between meningiomas and some benign or low-resolution gliomas. Despite this result, it indicates the model's good performance and lack of bias in discriminating for this group, as well as its generalizability even in complex diagnoses.

● Brain Tumor Categories

The model's accuracy was 93.5%, recall was 94.0%, and F1 93.7%. This indicates the proposed model's ability to correctly classify most cases with a low percentage of incorrect predictions. The results obtained in this category indicate the most distinctive morphological characteristics of the tissue.

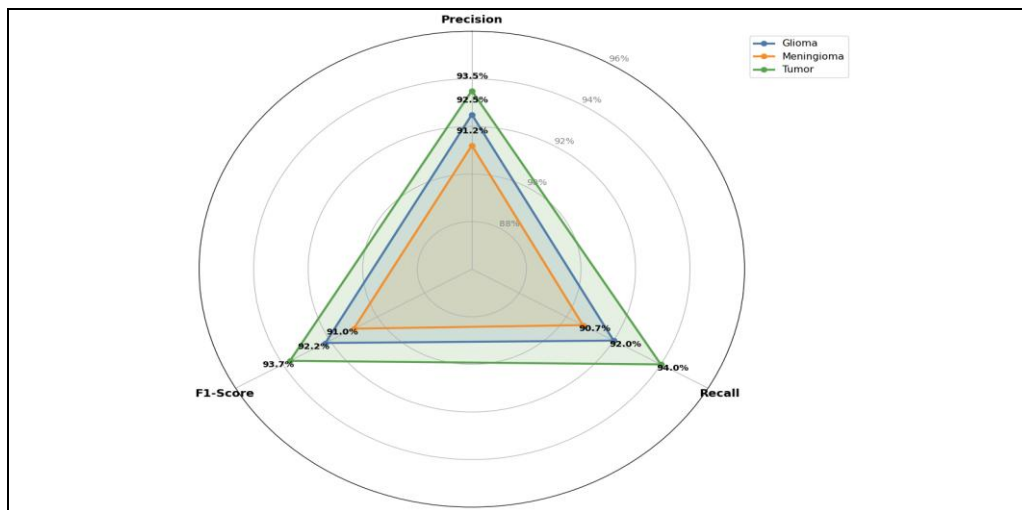


Figure 5: Radar diagram for model's performance Precision- Recall- F1-Score.
Source: Authors, (2026).

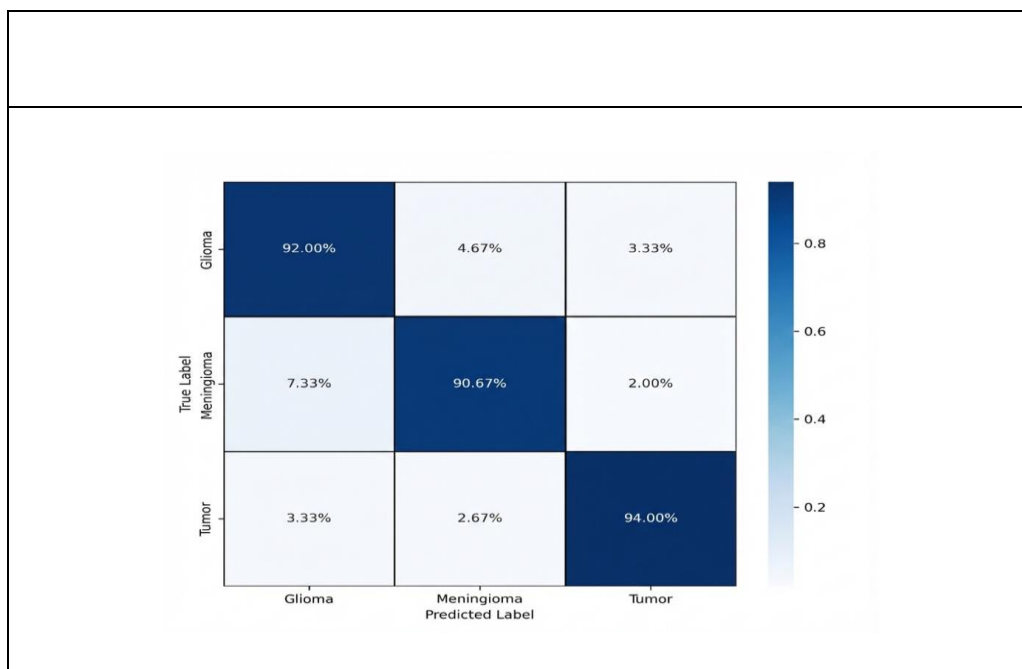


Figure 6: Confusion matrix with row-wise normalization, highlighting the sensitivity (recall) for the Glioma, Meningioma, and Tumor classes.

Source: Authors, (2026).

Figure 6, demonstrate Confusion matrix normalization, Glioma, Meningioma and Tumor classes Recall. A quantitative matrix characterization shows that this model is robust and consistent in the three categories. In the Glioma category, 138 images were correctly identified, with 7 misclassified as Meningioma and 5 as Tumor. In Meningioma, 136 cases were classified as the correct category, 11 misclassified as Glioma and 3 as Tumor. In Tumor, 141 of the cases were identified correctly, 5 misclassified as Glioma and 4 as Meningioma. These findings suggest that the model can distinguish well between the categories, with very low error rates. The spread of most confusion matrix values along the core diagonal suggests that there is excellent classification in all the categories. Among these, the most confusion was observed in regard to the Glioma and Meningioma segments, which is of clinical significance considering the obvious visual similarity of the histological structures on MRI images for these two tumor types. The results confirm that the Hybrid model is also an extremely good approach for brain tumor classification where the confusion matrix is almost as good as the actual model and can reach approximately 94% recall. This performance has a fair spread of prediction in all types and further increases its credibility with respect to AI-based medical diagnostic use cases.

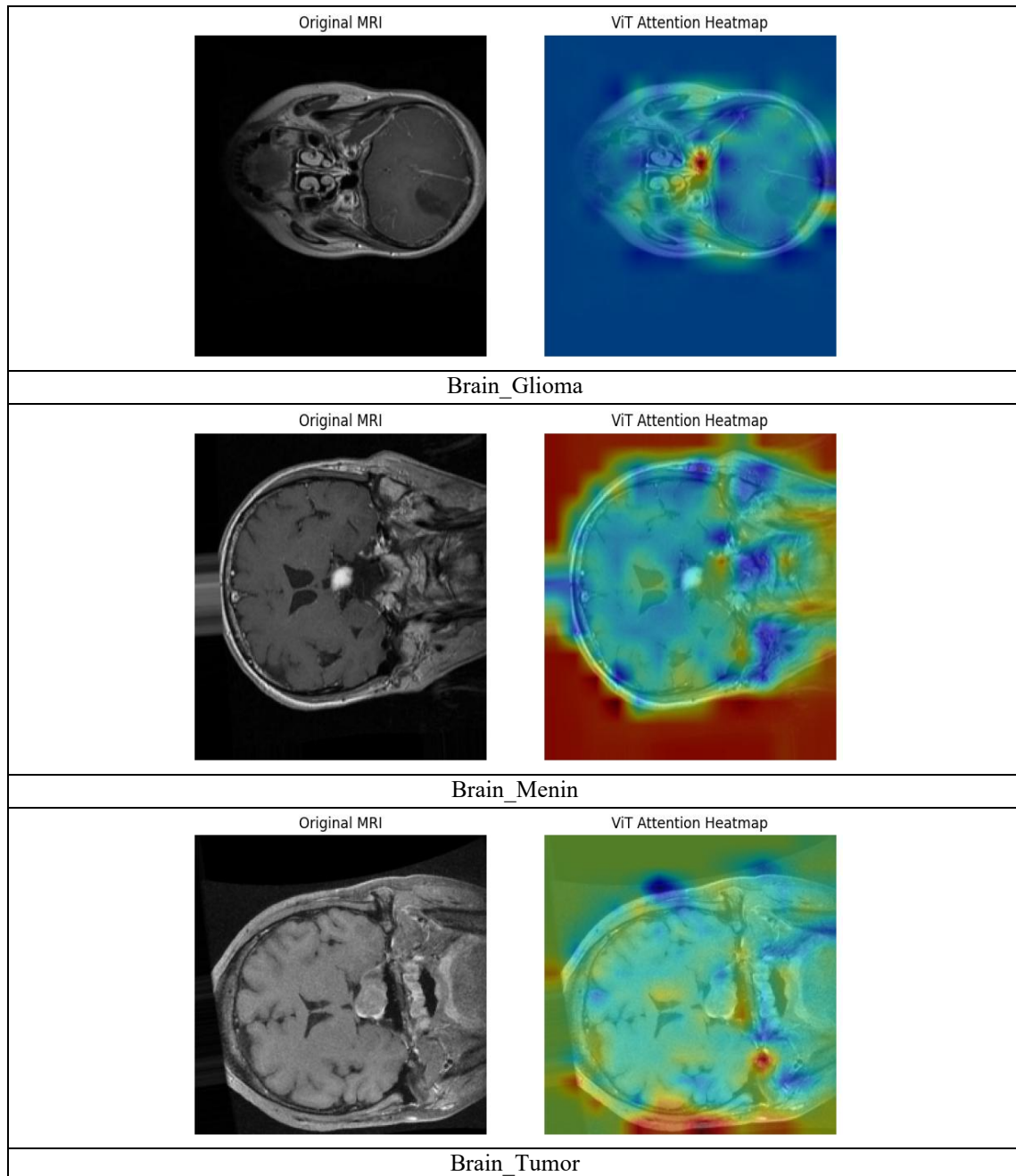


Figure 7: Attention maps from the ViT_BiLSTM Model on MRI images.

Source: Authors, (2026).

Figure 7, show that Attention maps obtained from the Vision Transformer model show that the highest attention values are concentrated in specific brain regions that anatomically correspond to tumor locations seen in reference MRI images. A distinct concentration of high-intensity attention areas (depicted by warm colors) was noted around the pathological lesions, contrasted by a marked decrease in attention values in the surrounding healthy tissue. According to the above findings, the model focuses primarily on regions related to classification, avoiding non-lesion-related features. Table.(1), Comparasion proposed methodology with other litruture.

Table.1: Comparasion FrameWork with Litruture.

Reference	Dataset	Model	Performance	Limintation
Agrawal et al [22]	DIDA handwritten digit string dataset	Decoded-ViT and CNN-LSTM	Character 90.30%, 85.31% Word 70.15%, 55.75%	Limited feature representation capacity and lack of attention-based global modeling
Safdar and Saadeldin [23]	MedMNIST Dataset Family.	CNN and ViT	84.5%	All experiments were conducted on small, simplified datasets (MedMNIST).No analysis of potential biases in attentional areas.
Dayan [24]	Chest X-ray images (COVID-19)	CNN and ViT	96.58%	Unable to achive equilibrium between each of accuracy and efficiency can lead to information loss through division.
Chen, et al [25]	Brain MRI	ViT-V-Net	0.726%	ViT is used only as a spatial encoder without temporal contextual analysis.
Jiang [26]	BC5CDR dataset	CNN, Bidirectional LSTM, Transformer	0.91 and 0.71	Limitations of generalization on high-resolution images; lack of explicit handling of domain shifts.
Remyal et al [27]	TORG0 and UA Speech	Hybrid CNN-BiLSTM-Transformer	98.74% and 99.86% for binary dysarthria detection, and 95.69% and 97.91% for multi-class severity	Sequential and contextual features. ViT handles features only in parallel. There is no modeling of the correlation between features.
Lu et al [28]	de-identified discharge summary data made available by Harvard University in 2008	CNN, Transformer encoder, a pretrained BERT, RNN , GRU , LSTM , Bi-LSTM, BERT-Base	91.3%-94.0%	Handles medical texts only. Not applicable to medical images. Does not handle MRIs or brain tumors.
OUR PROPOSED	Brain tumor (MRI dtatset)	Hybrid Vit with BiLSTM	93.0%	Integrated spatial-sequential representation leads to superior performance

Source: Authors, (2026).

V. CONCLUSIONS

The results of the experimental test show that the proposed hybrid approach, based on Vision Transformer (ViT) and Bidirectional Long Short-Term Memory Networks (Bi-LSTM), outperformed all other classes of brain tumors by getting better classification with respect to brain MRI images. This performance superiority is demonstrated with respect to the high values of performance metrics, especially overall accuracy and the mean F1 score across all categories and the balanced error distribution, as shown by the confusion matrix. The ambiguity matrix verifies that the majority of the samples were appropriately classified — with the most valuable values being aggregated within the main diagonal of the matrix, indicating the accuracy with which the model is able to separate and classify a variety of categories. There was only a marginal degree of ambiguity in classification of the glioma and meningioma, which is clinically expected because of the structural and histological close relation between these two categories in some instances. Conversely, the brain tumor category recorded high correct classification rate, indicating that the model worked effectively with the unique structural features in this class. Contrastingly, the accuracy and loss curves over the training and validation stages exhibited stability between train and validation. There was a tendency towards increasing validation accuracy with the loss values progressively decreasing, with no gap on this data between the training and test.

This tendency shows the good effect of the regularization model such as Dropout and Batch Normalization in controlling the model overfitting as well as generalization. The overall superior performance of the proposed model is attributed to the structural integration of the proposed components. Vision Transformers facilitate the learning of global context representations via a multi-headed self-attention mechanism, and so modeling long-range dependencies of medical images. Meanwhile, Bi-LSTM layers enhance the utilization of these representations by analyzing sequential correlations among features extracted. This way we can gain an even richer understanding among complex spatial relationships in MRI images. This integration is important to break out of the drawbacks of conventional convolutional techniques which mainly contain only the local features. Compared to the earlier works that applied single-pattern methods, present findings can be indicative of a striking effect on achieving a trade-off between accuracy and sensitivity for various classes. It also supports the possibility of hybrid architectures for brain tumor classification problems, especially when high heterogeneity and variability medical data is present.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Nagham Salim Mohammed, Omar S. Almolaa, Ahmed Saadi Abdullah, and Khalil I. Alsaif.
Methodology: Nagham Salim Mohammed and Khalil I. Alsaif.
Investigation: Nagham Salim Mohammed and Omar S. Almolaa.
Discussion of results: Nagham Salim Mohammed, Omar S. Almolaa and Khalil I. Alsaif.
Writing – Original Draft: Nagham Salim Mohammed, Omar S. Almolaa and Ahmed Saadi Abdullah.
Writing – Review and Editing: Nagham Salim Mohammed and Omar S. Almolaa Ahmed Saadi Abdullah.
Resources: Omar S. Almolaa.
Supervision: Nagham Salim Mohammed and Khalil I. Alsaif.
Approval of the final text: Nagham Salim Mohammed, Omar S. Almolaa, Ahmed Saadi Abdullah, and Khalil I. Alsaif.

VII. REFERENCES

- [1] Q. T. Ostrom, M. Price, C. Neff, G. Cioffi, K. A. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019," *Neuro-Oncology*, vol. 24, no. S5, pp. v1–v95, 2022.
- [2] Y. Cakmak and I. Pacal, "Comparative analysis of transformer architectures for brain tumor classification," *Exploration of Medicine*, 2025.
- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition Cakmak and I. Pacal, "Comparative analysis of transformer architectures for brain tumor classification," *Exploration of Medicine*, 2025.
- [3] H. Ahmed, M. O. Dada, and B. Samaila, "Current challenges of the state-of-the-art of AI techniques for diagnosing brain tumor," *Material Science & Engineering*, vol. 7, no. 4, pp. 196–208, 2023.
- [4] N. Ullah, A. Javed, A. Alhazmi, S. M. Hasnain, A. Tahir, and R. Ashraf, "TumorDetNet: A unified deep learning model for brain tumor detection and classification," *PLoS ONE*, vol. 18, no. 9, p. e0291200, 2023.
- [5] O. Shobayo and R. Saatchi, "Developments in deep learning artificial neural network techniques for medical image analysis and interpretation," *Diagnostics*, vol. 15, no. 9, p. 1072, 2025.
- [6] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, p. 1273253, 2023.
- [7] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in scale*, arXiv preprint arXiv:2010.11929, 2020.
- [9] S. H. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, Art. no. 200, 2022.
- [10] Z. Pei, Y. Gou, M. Ma, M. Guo, C. Leng, Y. Chen, and J. Li, "Alzheimer's disease diagnosis based on long-range dependency mechanism using convolutional neural network," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 36053–36068, 2022.
- [11] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual attention vision transformers," arXiv preprint arXiv:2204.03645, 2022.
- [12] M. A. Talukder, M. M. Islam, M. A. Uddin, M. A. Layek, U. K. Acharjee, and M. A. Moni, "A deep ensemble learning framework for brain tumor classification using data balancing and fine-tuning," *Scientific Reports*, vol. 15, Art. no. 35251, 2025.
- [13] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac, K. Khezeli, and P. Rashidi, "Transformers and large language models in healthcare: A review," *Artificial Intelligence in Medicine*, vol. 154, p. 102900, 2024.
- [14] R. Kumar and S. Kumar, "Effectiveness of vision transformers in human activity recognition from videos," in *Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 2023, pp. 1–6, doi: 10.1109/InCACCT57535.2023.10141761.
- [15] M. Aly, A. Ghallab, and I. S. Fathi, "Tumor ViT-GRU-XAI: Advanced brain tumor diagnosis framework: Vision transformer and GRU integration for improved MRI analysis: A case study of Egypt," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3513235.
- [16] S. Tummla, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling," *Current Oncology*, vol. 29, no. 10, pp. 7498–7511, 2022, doi: 10.3390/curroncol29100590.
- [17] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, and M. Islam, "Combining the transformer and convolution for effective brain tumor classification using MRI images," *Applied Sciences*, vol. 13, no. 6, p. 3680, 2023, doi: 10.3390/app13063680.
- [18] S. Mehta and D. Kundra, "A robust CNN-LSTM framework for sensitive and specific brain tumor classifications," in *Proceedings of the International Conference on Automation and Computational Intelligence*, 2025.
- [19] S. R. Rekha, B. Karthik, and T. Vijayan, "Developing a hybrid framework for brain tumor multi-classification from MRI images using LSTM," in *Proceedings of IEEE INDIACom 2025*, 2025, pp. 1–6.
- [20] N. L. Mohammed, M. S. Aziz, and O. I. AlSaif, "Design and implementation of robot control system for multistory buildings," *Telkomnika Telecommunication Computing Electronics and Control*, vol. 18, no. 5, pp. 2682–2689, 2020.
- [21] K. M. Sujon, K. Choi, and M. A. Samad, "Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *Journal of Big Data*, vol. 12, p. 268, 2025.
- [22] V. Agrawal, J. Jagtap, and M. V. P. Kantipudi, "Decoded-ViT: A vision transformer framework for handwritten digit string recognition," *Revue d'Intelligence Artificielle*, vol. 38, no. 2, pp. 523–529, 2024.

- [23] A. Safdar and M. Saadeldin, "CoMViT: An efficient vision backbone for supervised classification in medical imaging," University College Dublin, 2025.
- [24] B. Dayan, "Lung disease detection with vision transformers: A comparative study of machine learning methods," National University of Mongolia, 2024.
- [25] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration," arXiv preprint arXiv:2104.06468, 2021.
- [26] S. Jiang, "Medical named entity recognition based on bidirectional gated pyramid network," in Proceedings of the 3rd International Conference on Software Engineering and Machine Learning, 2025.
- [27] M. S. Remya, P. Ishwar, and P. Nedungadi, "A hybrid cross-attentive CNN-BiLSTM-transformer network for dysarthria severity classification," Scientific Reports, 2025.
- [28] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," BMC Medical Research Methodology, 2022.