International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

Attention-Guided Hybrid Deep Learning Framework for Explainable Brain Tumour Classification Using MRI

P.Sri Muthu Bharathi[1], Mrs.J.Sanetha Baegam[2]

[1]PG Scholar Bio-Medical Department, Udaya School of engineering, Kanyakumari, Tamil Nadu, India.
[2]Assistant Professor Bio-Medical Department, Udaya School of engineering, Kanyakumari, Tamil Nadu, India.

Abstract-*Classifying brain tumors accurately from MRI scans is a critical but challenging task in medical imaging. Traditional deep learning models like VGG16, ResNet50, and EfficientNet have made significant strides in this field. However, these models sometimes struggle to distinguish between similar tumor types such as Glioma and Meningioma, partly because they can focus on irrelevant parts of the brain rather than the tumor itself. To tackle this issue, researchers have developed a new approach called the Attention-Guided Tumour-Focused Convolutional Neural Network (AGT-CNN). This innovative framework starts by using a U-Net segmentation model to isolate the brain and tumor regions from MRI images. By removing background noise and unrelated anatomical details, the model can focus more precisely on the areas that matter. Next, the isolated regions are analyzed using a hybrid feature extractor that combines the strengths of VGG16 and ResNet50. This is enhanced with special attention mechanisms that highlight the tumor-relevant features both spatially and across channels, helping the model to concentrate on the most important information. Training techniques such as early stopping are used to ensure the model learns effectively without over fitting. Beyond accuracy, the study also addresses the interpretability of the AI system. To build trust and transparency, the researchers apply explain ability tools like Grad-CAM and LIME alongside the model's own attention maps. These methods reveal how the model makes decisions, showing that it truly focuses on tumor regions rather than irrelevant brain structures. This work underscores the value of combining region-focused preprocessing with attention-guided learning and multi-stage explain ability. Together, these elements help create AI tools for brain tumor classification that are not only powerful but also interpretable and trustworthy, paving the way for more reliable medical diagnoses.*

Key Words: AGT-CNN, U-Net Segmentation, Hybrid Feature Extraction, Explainable AI, Grad-CAM, LIME

1. INTRODUCTION

Brain tumor classification using MRI has become a vital area of medical research because early and accurate diagnosis can save lives. Thanks to advances in imaging technology and artificial intelligence, automated systems now help doctors detect and classify tumors more reliably, reducing human error and improving decision-making [1]. Deep learning, especially convolutional neural networks (CNNs), has shown great success in analyzing MRI scans. Models like VGG and ResNet are popular because they can extract detailed features from images, which helps identify and segment tumors [3], [4]. However, these models sometimes get distracted by irrelevant background details and struggle to clearly distinguish between tumor types. To overcome this, researchers introduced attention mechanisms that help models focus on the most important parts of the images [5]. Besides classifying tumors, accurately locating them is crucial. Benchmark datasets like BRATS provide multimodal MRI data that support the creation of powerful segmentation models [8]. Preprocessing techniques, such as bias field correction, also improve image quality and help models learn better [9]. Another key challenge is making deep learning models interpretable. In medicine, it's important to understand why a model makes a specific prediction to build trust. Explainable AI methods like Grad-CAM and LIME offer visual explanations by highlighting the brain regions influencing the decisions [6], [7], ensuring that the model's focus is on relevant tumor areas. Despite progress, challenges remain. Tumors vary widely in size, shape, and intensity, making it hard for models to generalize. MRI scans often contain noise and complex anatomy that can confuse algorithms. Segmentation approaches, especially using architectures like U-Net, help isolate tumor regions and reduce background distractions, improving classification accuracy [2]. Hybrid deep learning models that combine architectures like VGG and ResNet offer promising results by leveraging their complementary strengths [3], [4]. When these models are combined with attention mechanisms and explainable AI, they not only boost accuracy but also provide meaningful insights into

their predictions [5], [6]. Such advances are essential for deploying reliable and transparent AI diagnostic tools in clinical practice. Overall, recent research focuses on creating unified frameworks that integrate segmentation, classification, and explainability into a seamless system [2], [10]. These integrated approaches enhance both performance and transparency, making AI tools more suitable for real-world clinical use. Ultimately, these innovations hold great promise for improving early detection, diagnosis, and treatment planning for brain tumor patients, leading to better healthcare outcomes. In addition, the growing availability of large-scale medical imaging datasets and increased computational power has significantly accelerated research in brain tumor analysis. Modern deep learning frameworks enable faster training and deployment of complex models, making real-time diagnosis more feasible. However, challenges such as data imbalance, limited labeled datasets, and variability across imaging protocols still need to be addressed for consistent performance across diverse clinical settings [8], [10]. Therefore, developing robust, scalable, and generalizable models remains a key focus in advancing AI-driven brain tumor classification systems.

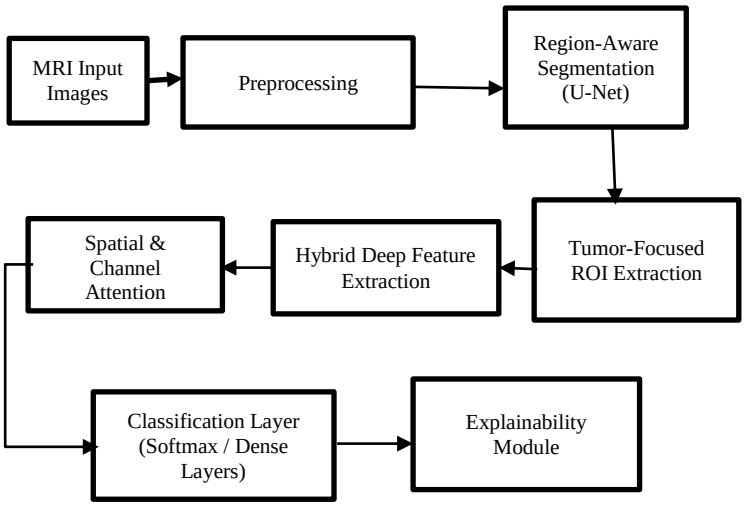

Fig 1: Block Diagram

2. METHODOLOGIES

In the quest to improve brain tumor classification, researchers have designed a smart system that begins with MRI images. These images first go through preprocessing to sharpen details and clear away any noise, setting the stage for accurate analysis. Next, a sophisticated segmentation step uses a U-Net-based module to precisely separate the brain and tumor areas, ensuring the system focuses on what truly matters. From here, the system extracts regions of interest (ROI) centered on the tumor, filtering out irrelevant parts. These focused areas are then analyzed by a hybrid deep learning module that blends several convolutional neural networks (CNNs). This combination helps capture a rich mix of subtle and complex features, making the model more effective at recognizing different tumor types.To boost accuracy further, the system applies spatial and channel attention mechanisms. These help the model zero in on the most critical tumor-specific patterns, refining its understanding even more. The enhanced features are then passed to a classification layer that predicts the tumor category, typically using softmax or dense neural layers. What sets this system apart is its integrated explain ability module. This component offers visual, easy-to-understand insights into how the model makes decisions, helping doctors trust and interpret the results more confidently. By combining precise segmentation, attention-driven feature learning, and transparent explanations, the system offers a balanced, reliable, and interpretable approach to brain tumor classification. Overall, this architecture creates a smooth, efficient pipeline where each stage builds on the last, reducing distractions from irrelevant details and strengthening tumor-specific analysis. It not only improves performance but also supports medical professionals in understanding and trusting the technology behind their critical diagnoses.

2.1 MRI Data Acquisition & Preprocessing Module

The MRI Data Acquisition and Preprocessing Module gets ready for accurate brain tumor analysis by creating good quality images. It starts by collecting different types of MRI scans like T1, T2, FLAIR, and contrast-enhanced images. These scans give extra details about the brain tissues. To make sure all the images are the same, they are adjusted for brightness and size. This helps the models used in deep learning work better. Next, steps like reducing noise and improving contrast are done to make the images clearer and show the tumor areas better.

Also, techniques like rotating, flipping, and resizing the images are used to make the dataset more varied. This helps prevent the model from learning too much from one type of image. In the end, this module makes sure the MRI images are all the same, clean, and improved. This helps in better classification of tumors.

2.2 Region-Aware Tumour Segmentation Module (U-Net)

The Region-Aware Tumour Segmentation Module is an important part of the system that helps accurately

separate tumour areas from brain MRI images using a U-Net-based design. This step ensures that only the parts of the image that are related to tumours are sent to the next stage, which helps make the results more accurate and less affected by other things in the image. The segmentation process works with an encoder-decoder U-Net structure that includes skip connections. This setup lets the model learn both the general context and the specific details of the brain images. The encoder gradually picks up high-level information from the MRI slices, understanding complex patterns and how different parts of the brain relate to each other. This helps the model recognize the structure and position of possible tumours. The decoder then rebuilds the segmented image by increasing the resolution of the encoded data, focusing on creating clear boundaries around the tumours. This ensures the model can accurately locate tumours, even if they are not clearly defined or spread out. A key advantage of the U-Net structure is the use of skip connections, which pass detailed spatial information from the encoder to the decoder. These connections keep important details about the anatomy that could otherwise be lost during the process of reducing image size, leading to more accurate and detailed results. The module creates detailed maps that show exactly where the tumours are in the brain. By focusing on the tumour areas, the system removes unwanted parts like the skull and background noise. This segmentation helps improve the performance of the next steps in the system, as the model can concentrate on the features that are specific to tumours rather than other unnecessary details. Overall, this module is essential for making the brain tumour analysis system more accurate and dependable.

2.3 Tumour-Focused ROI Extraction

The Tumour-Focused Region of Interest (ROI) Extraction Module is designed to refine the segmented output by isolating only the tumour-specific regions from MRI images. This module plays a crucial role in ensuring that the subsequent classification model focuses exclusively on relevant tumour features, thereby improving accuracy and computational efficiency. In this stage, the segmentation masks generated from the U-Net module are applied to the original MRI images to precisely extract tumour regions. By using these masks, the system effectively isolates tumour-only areas, eliminating surrounding healthy tissues and irrelevant background information. The module then performs cropping or masking operations to remove non-essential regions, ensuring that the input to the CNN contains only tumour-centered data. This not only improves the quality of the input but also reduces input dimensionality, leading to lower computational requirements and faster processing. By focusing on tumour regions, the system significantly improves the signal-to-noise ratio, allowing the model to learn more meaningful and discriminative features. This helps in enhancing classification performance, especially in challenging cases where tumour boundaries are subtle or complex. Additionally, this module helps prevent CNN models from learning spurious correlations that may arise from irrelevant anatomical structures. By standardizing inputs to contain only tumour regions, it ensures consistency across samples, which is essential for stable and reliable model training. Furthermore, the tumour-focused inputs facilitate more effective attention learning, enabling attention mechanisms to concentrate on clinically significant regions without distraction. Overall, this module strengthens the robustness, efficiency, and accuracy of the proposed brain tumour classification system.

2.3 Hybrid Deep Feature Extraction Module (VGG16 + ResNet50)

The Hybrid Deep Feature Extraction Module is a key part of the system.

It is designed to get detailed and useful information about tumor areas by combining the best parts of two strong deep learning models: VGG16 and ResNet50. Using both models together helps the system understand both small details and bigger meanings in the images, which makes the classification better. VGG16 is used to get low-level and mid-level features like edges, textures, and simple patterns found in tumor regions. Because VGG16 has a simple and consistent structure with many layers, it is very good at picking up small changes in brightness and structure, which are important for finding the edges and different parts inside tumors. At the same time, ResNet50 is used to get deeper and more meaningful features. ResNet50 uses special connections called skip connections, which help the network learn more complex features without problems with learning over time. These connections make it easier to train very deep networks and help capture detailed tumor characteristics like shape, size, and how parts of the brain relate to each other. A major benefit of this module is combining the features from both VGG16 and ResNet50. The features are either put together directly or merged using special techniques, creating a full and detailed feature picture that uses the best parts of both networks. This combined feature space helps the model tell apart similar tumor types like glioma and meningioma, which often have similar looks. By combining structural details from VGG16 and contextual information from ResNet50, the module allows the model to learn more effectively

from different aspects of the data, leading to more reliable and accurate results. This approach also helps the model work well with different kinds of data and imaging conditions. Compared to models that use just one network, the hybrid design is more stable, gives a wider variety of features, and improves the ability to classify tumors. This makes it especially good for complex tasks like analyzing brain tumors. Overall, this module is very important for making the system more accurate and dependable.

2.4 Spatial & Channel Attention Module

The Spatial and Channel Attention Module helps the model focus better on important features related to tumors in the feature maps. Spatial attention finds where the important tumor areas are by highlighting those regions and ignoring the background. At the same time, channel attention finds out which features are most important by giving more weight to the useful ones, which makes the features better represented.

By using both these methods together, the module makes sure the important tumor patterns stand out and less important information is reduced. This leads to better accuracy in classification and helps tell apart similar types of tumors. Also, attention maps show which parts the model is looking at, making the process easier to understand. Overall, this module makes feature selection more effective, improves model performance, and increases transparency, making the system more trustworthy for use in clinical settings.

2.5 Classification & Optimization Module

The Classification and Optimization Module is the last part of the system, and its job is to predict the type of brain tumor using features that have been improved with attention mechanisms. Once the feature maps are created, they are flattened and sent through fully connected layers, which help the model learn how different features relate to each other and connect them to the final outputs. A Softmax layer is used to classify the tumor into multiple types, giving a probability score for each possible type. The model is trained using a loss function called cross-entropy, which helps reduce the errors in predictions. To make the model work well and avoid memorizing the training data, methods like early stopping, adjusting the learning rate, dropout, and regularization are used. These methods help the model perform better on new data and reach a stable solution faster. In total, this module ensures that the system can accurately, reliably, and consistently classify brain tumors, making it useful for real-life medical use.

2.6 Multi-Stage Explainability & Decision Interpretation

The Multi-Stage Explain ability and Decision Interpretation Module makes sure the deep learning model is clear and easy to understand, which makes it more trustworthy for use in healthcare. It uses attention maps to show which parts of the image are important for identifying tumors, giving built-in information about where the model looks when making a prediction. To make the model even easier to understand, methods like Grad-CAM and LIME are used. Grad-CAM creates heat maps that show which areas of the image affect the prediction. LIME breaks down the image into smaller parts and explains how each part contributes to the model's decision. Together, these tools help people better understand and check the model's decisions. Overall, this module builds trust, helps doctors verify the model's results, and ensures the predictions are both correct and easy to explain, making the system ready for use in actual medical settings**.**

RESULT & DISCUSSION

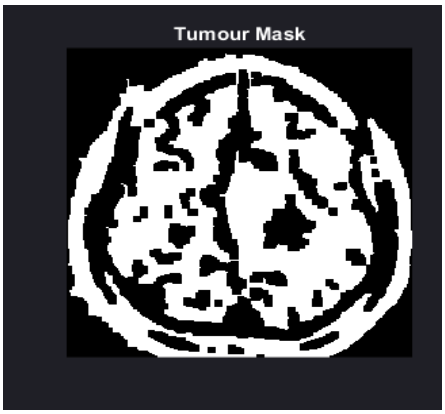

Fig.2. Segmentation Result

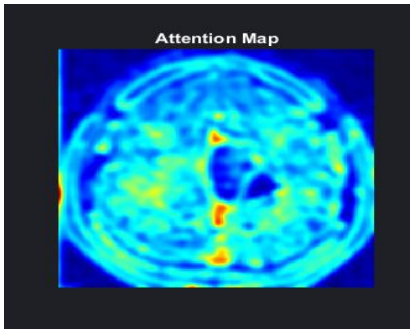

Fig 3: Attention Map

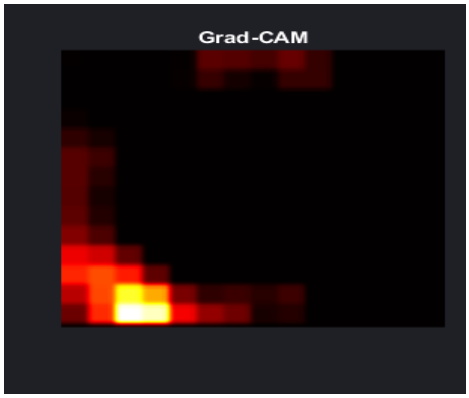

Fig 4: Gradient-weighted Class Activation Map

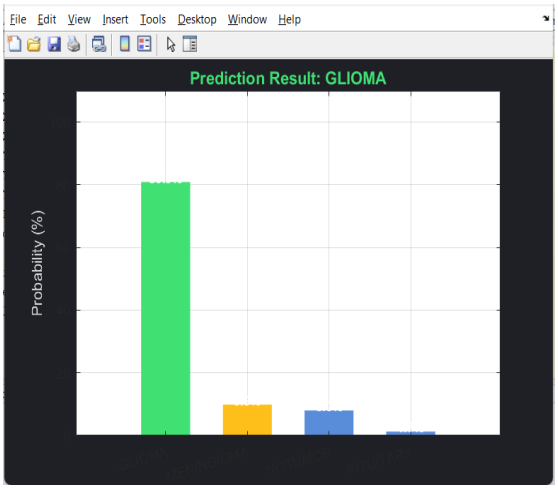

Fig 5: Prediction Result

The experimental outcomes showcase the effectiveness of the proposed AGT-CNN model in accurately distinguishing brain tumors from MRI scans. Fig: 2 displays the segmentation results achieved with the U-Net model. The tumor areas are distinctly separated from the surrounding brain tissue, showing that the segmentation module efficiently removes irrelevant structures such as the skull and background. This enhances the quality of the input data for later stages and ensures the model focuses solely on clinically relevant areas. Fig: 3 depicts the attention map resulting from the spatial and channel attention mechanisms. The highlighted regions indicate that the model successfully identifies and prioritizes tumor-specific areas within the feature maps. This confirms that the attention module improves feature learning by emphasizing relevant details and neglecting irrelevant ones. Fig: 4 presents the Grad-CAM visualization, offering a heatmap of the regions most impactful for the model's prediction. The high-intensity regions concentrated over the tumor area confirm that the model bases its decisions on meaningful features. This supports the interpretability of the proposed system and ensures predictions are not skewed by irrelevant areas. Fig: 5 shows the final prediction outcome, where the model identifies the tumor type alongside probability scores.

The high confidence in predictions suggests the model has effectively learned discriminative features for accurate classification. This outcome validates the robustness and reliability of the proposed method. Overall, the results indicate that integrating segmentation, attention mechanisms, and explain ability techniques significantly enhances classification accuracy and model transparency. The system not only provides accurate predictions but also offers visual explanations, making it suitable for real-world clinical use.

CONCLUSION

This work introduces an Attention-Guided Tumour-Focused CNN (AGT-CNN) for precise brain tumour classification using MRI scans. The framework combines U-Net-based segmentation, a hybrid feature extraction approach (VGG16 + ResNet50), and attention mechanisms to improve tumour-specific learning and minimize misclassification, especially among visually similar tumour types. The integration of explain ability methods such as attention maps, Grad-CAM, and LIME enhances model transparency by offering clear visual insights into the model's predictions. This ensures that the model's decisions are grounded in relevant tumour regions, thereby boosting trust and practicality in clinical environments. In summary, the proposed system achieves enhanced accuracy, robustness, and interpretability. It shows significant potential for real-world medical applications by enabling reliable and efficient brain tumour diagnosis. Future research could focus on extending the model to larger and more varied datasets and optimizing it for real-time clinical use.

REFERENCES

[1] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1240–1251, May 2016.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations, 2015.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[6] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[8] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[9] N. J. Tustison et al., "N4ITK: Improved N3 bias correction for MRI," IEEE Transactions on Medical Imaging, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.

[10] H. Chaddad, M. Desrosiers, and M. Toews, "Deep radiomic analysis of MRI related to Alzheimer's disease," IEEE Access, vol. 6, pp. 58213–58221, 2018.