International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

NeuroXAI-Net: An Explainable Ensemble Transfer Learning Architecture for Multiclass Brain Tumour Classification from MRI Scans

Mrs. M. Sujana Priyadarshini[1], Vinnakoti Sakyavardhan[2]
[1]Associate Professor, [2]M. tech Student,
Department of Computer Science & Engineering(AI), Pydah College of Engineering,
Yanam Road, Tallarevu, Patavala, Andhra Pradesh, 533461

Abstract- Brain tumour diagnosis using Magnetic Resonance Imaging (MRI) plays a crucial role in early treatment planning and patient survival. However, manual interpretation of MRI scans is time-consuming and may lead to inconsistent clinical decisions. To address these limitations, this study proposes an explainable ensemble transfer learning framework for multiclass brain tumour classification. The proposed model integrates multiple pre-trained convolutional neural network architectures and aggregates their predictions using an ensemble strategy to enhance classification robustness and reduce overfitting. Furthermore, Explainable Artificial Intelligence (XAI) techniques are incorporated to visualize tumour regions and improve model interpretability, thereby increasing clinical trust and reliability. The dataset consists of multiclass MRI images categorized into glioma, meningioma, pituitary tumour, and no-tumour classes. Data augmentation and preprocessing techniques are employed to improve generalization performance. Experimental evaluation demonstrates that the ensemble framework achieves superior classification accuracy compared to individual transfer learning models. Performance is assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis. The integration of explainability tools further validates the model's capability to focus on clinically relevant tumour regions. The proposed approach offers a reliable, scalable, and interpretable solution for automated brain tumour detection and classification, making it suitable for real-world clinical decision support systems.

Keywords – Brain tumour classification, Transfer learning, Ensemble learning, Vision transformers, Deep learning, MRI analysis, Explainable AI (XAI), Multiclass classification, medical image analysis, Convolutional neural networks.

I. INTRODUCTION

Brain tumours represent one of the most critical neurological disorders, characterized by abnormal and uncontrolled cell growth within brain tissues. These tumours can be broadly categorized into primary tumours, which originate in the brain, and secondary (metastatic) tumours, which spread from other parts of the body. Early detection and accurate classification of brain tumours are essential for effective treatment planning, prognosis estimation, and improving patient survival rates. Magnetic Resonance Imaging (MRI) is widely regarded as the most reliable and non-invasive imaging modality for detecting and analysing brain tumours due to its superior soft-tissue contrast and high-resolution imaging capabilities. Traditionally, radiologists manually examine MRI scans to identify tumour presence and type. However, manual interpretation is time-consuming, subjective, and may result in inter-observer variability. Furthermore, the increasing volume of medical imaging data has made manual analysis increasingly challenging. These limitations have motivated the development of automated computer-aided diagnosis (CAD) systems using machine learning (ML) and deep learning (DL) techniques. In recent years, convolutional neural networks (CNNs) and transfer learning (TL) approaches have demonstrated significant success in medical image classification tasks. Transfer learning enables pre-trained deep neural networks to adapt to medical imaging datasets, thereby reducing computational requirements and improving convergence speed. However, single deep learning models may suffer from overfitting, limited generalization capability, and reduced robustness when trained on relatively small medical datasets.

To overcome these challenges, ensemble learning techniques have been introduced to combine the strengths of multiple models. By aggregating predictions from different transfer learning architectures, ensemble models enhance classification stability, improve feature representation, and reduce prediction bias. Additionally, despite their high performance, deep learning models are often criticized as "black-box" systems due to their lack of interpretability. In clinical applications, model transparency is essential to gain trust from medical practitioners and ensure regulatory compliance. Therefore, this

research proposes an explainable ensemble transfer learning framework for multiclass brain tumour classification using MRI images. The proposed approach integrates multiple pre-trained convolutional neural network architectures and employs ensemble strategies to improve accuracy and robustness. Furthermore, Explainable Artificial Intelligence (XAI) techniques are incorporated to visualize tumour regions and validate the model's decision-making process.

The main contributions of this work are summarized as follows:

- Development of a multiclass brain tumour classification system using transfer learning-based deep neural networks.
- Design of an ensemble framework to enhance prediction accuracy and reduce overfitting.
- Integration of Explainable AI techniques to improve model interpretability and clinical reliability.
- Comprehensive performance evaluation using accuracy, precision, recall, F1-score, and confusion matrix analysis.

The remainder of this paper is organized as follows. Section II presents the related work and literature review. Section III describes the proposed methodology and system architecture. Section IV discusses performance evaluation metrics. Section V presents experimental results and discussion. Finally, Section VI concludes the paper and outlines future research directions.

II. LITERATURE SURVEY

Automated brain tumour detection and classification have gained significant attention in recent years due to advancements in machine learning and deep learning techniques. Researchers have explored various approaches ranging from traditional machine learning algorithms to advanced deep neural network architectures for medical image analysis. Early research focused on conventional machine learning techniques combined with handcrafted feature extraction methods. Techniques such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest, and Naïve Bayes were widely used for tumour classification tasks. These methods relied heavily on manual feature engineering, including texture features, histogram-based features, and wavelet transforms extracted from MRI images. Although these approaches achieved moderate classification accuracy, their performance was limited by feature dependency and reduced generalization capability.

With the emergence of deep learning, Convolutional Neural Networks (CNNs) significantly improved medical image classification performance. CNN-based architectures automatically learn hierarchical feature representations directly from raw MRI images, eliminating the need for manual feature extraction. Several researchers have applied architectures such as VGGNet, ResNet, DenseNet, and Inception networks for binary and multiclass brain tumour classification. These models demonstrated improved accuracy; however, they often required large annotated datasets and high computational resources. Transfer learning has been widely adopted to address the challenge of limited medical imaging datasets. By utilizing pre-trained models trained on large-scale datasets, researchers fine-tuned network parameters for brain tumour classification tasks. Transfer learning not only reduces training time but also improves convergence stability and overall accuracy. Despite these advantages, single transfer learning models may suffer from overfitting, especially when trained on small or imbalanced medical datasets.

To enhance robustness and prediction reliability, ensemble learning approaches have been introduced. Ensemble methods combine predictions from multiple models to reduce variance and improve generalization performance. Studies have shown that combining CNN architectures through majority voting, weighted averaging, or stacking techniques results in improved classification accuracy compared to individual models. However, ensemble methods increase model complexity and computational cost. Recently, Vision Transformers (ViT) have emerged as a powerful alternative to traditional CNNs in image classification tasks. Unlike CNNs, Vision Transformers capture global contextual relationships using self-attention mechanisms. Several studies have demonstrated that transformer-based architectures achieve competitive performance in medical image analysis. Nevertheless, their application in brain tumour classification remains relatively limited, particularly when integrated with ensemble strategies.

Another critical concern in medical AI systems is interpretability. Deep learning models are often criticized as "black-box" systems, making it difficult for clinicians to understand the reasoning behind predictions. To address this issue, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, SHAP, and saliency maps have been incorporated into tumour detection frameworks. These techniques provide visual explanations by highlighting important tumour regions, thereby improving clinical trust and transparency. Despite significant progress, existing studies face challenges such as dataset imbalance, limited interpretability, model overfitting, and computational inefficiency. Moreover, few studies integrate transfer learning, ensemble modelling, and explainability into a unified framework for multiclass brain tumour classification. Therefore, this research aims to bridge these gaps by proposing an explainable ensemble transfer learning framework that enhances classification accuracy, improves generalization capability, and provides interpretable decision support for clinical applications.

III. SYSTEM ANALYSIS

A. Existing System

Existing brain tumour detection systems primarily rely on conventional machine learning and standalone deep learning models for classification of MRI images. Early approaches utilized traditional algorithms such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Trees, Random Forest, Naïve Bayes, and Logistic Regression. These methods required manual feature extraction techniques including texture analysis, histogram features, and wavelet transforms before classification. With the advancement of deep learning, Convolutional Neural Networks (CNNs) became widely adopted for automatic feature extraction and tumour classification.

Several studies implemented pre-trained architectures such as VGGNet, ResNet, DenseNet, and Inception models through transfer learning. These approaches significantly improved classification accuracy compared to traditional machine learning techniques. Some research works further explored Vision Transformers (ViT) to capture global contextual relationships within MRI images. However, most existing systems rely on single-model architectures. Although these models achieve good performance, they often lack robustness and may not generalize well across diverse datasets. Additionally, many systems focus only on prediction accuracy without providing interpretability or clinical justification for model decisions.

DISADVANTAGES OF THE EXISTING SYSTEM

- **Lack of Interpretability**:

Most deep learning models operate as black-box systems. In medical diagnosis, it is crucial to understand how the model arrives at a decision. Without explainability, clinical adoption becomes limited.

- **Overfitting and Limited Generalization:**

Single-model deep learning architectures may overfit when trained on small or imbalanced MRI datasets. This reduces their performance on unseen data.

- **Dataset Imbalance Issues:**

Medical imaging datasets often contain class imbalance among tumor types, which affects classification reliability and prediction fairness.

- **High Computational Complexity:**

Deep neural networks, especially transformer-based models, require significant computational resources, making deployment challenging in resource-constrained environments.

- **Absence of Model Robustness:**

Standalone models are more sensitive to noise and minor variations in image quality, which can affect diagnostic consistency.

- **Limited Clinical Validation:**

Many systems focus on accuracy metrics without providing visualization or localization of tumour regions, reducing trust among healthcare professionals.

B. Proposed System

To overcome the limitations of existing approaches, this research proposes an Explainable Ensemble Transfer Learning Framework for multiclass brain tumour classification using MRI images. In the proposed system, MRI images are first subjected to preprocessing steps including resizing, normalization, augmentation, and noise reduction to improve data quality and model generalization. The dataset is then divided into training, validation, and testing sets to ensure unbiased evaluation. Multiple pre-trained deep learning models such as CNN-based architectures and transformer-based models are fine-tuned using transfer learning.

Instead of relying on a single architecture, an ensemble strategy is implemented to combine predictions from multiple models using techniques such as majority voting or weighted averaging. This enhances classification robustness and reduces variance. To improve hyperparameter selection and optimize performance, systematic tuning techniques such as cross-validation are employed. The proposed system is evaluated using comprehensive performance metrics including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC analysis. Furthermore, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM are integrated to visualize tumour regions responsible for predictions. This improves transparency, interpretability, and clinical reliability of the model.

Advantages of the Proposed System

- Improved classification accuracy through ensemble learning.
- Reduced overfitting and enhanced generalization capability.
- Enhanced interpretability using explainability techniques.
- Better handling of multiclass tumour classification.
- Increased clinical trust and decision support reliability.

IV. SYSTEM DESIGN

SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture.

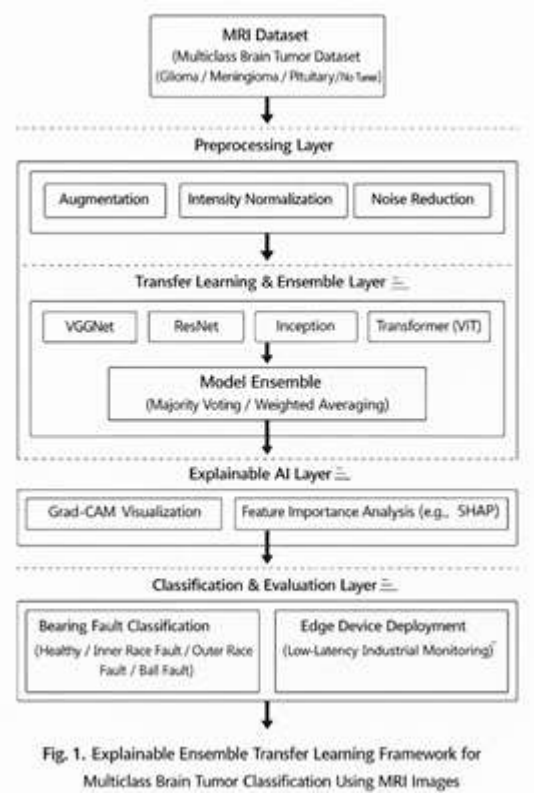

Fig. 1. Explainable Ensemble Transfer Learning Framework for Multiclass Brain Tumor Classification Using MRI Images

Fig 1. Methodology followed for proposed model

III. SYSTEM ANALYSIS

A. Existing System

MRI Data Acquisition and Preprocessing: The proposed framework begins with the collection of multiclass brain tumour MRI datasets comprising Glioma, Meningioma, Pituitary tumour, and No-Tumour classes. Since medical imaging data often contain variations in resolution, intensity distribution, and noise artifacts, a comprehensive preprocessing pipeline is implemented. Preprocessing steps include image resizing to a fixed input dimension, intensity normalization to ensure consistent pixel distribution, and noise suppression using filtering techniques. Data augmentation strategies such as rotation, flipping, zooming, and contrast adjustment are applied to enhance model generalization and mitigate overfitting. This stage ensures improved data quality and robustness prior to model training.

Transfer Learning-Based Feature Extraction: Instead of training models from scratch, the framework leverages pre-trained deep learning architectures for feature extraction. Convolutional Neural Networks (CNNs) such as VGGNet, ResNet, and Inception, along with Vision Transformer (ViT) models, are fine-tuned on the MRI dataset. These pre-trained models capture high-level spatial and structural patterns from MRI images, significantly reducing training time while improving classification performance. The extracted deep features serve as discriminative representations for tumour categorization.

Ensemble Learning Mechanism: To enhance predictive stability and reduce model variance, an ensemble strategy is adopted. Predictions obtained from individual transfer learning models are combined using majority voting and weighted averaging techniques. This ensemble mechanism improves classification reliability by minimizing individual model biases and enhancing generalization capability across tumor classes.

Explainable AI (XAI) Integration: Interpretability is critical in medical diagnostic systems. Therefore, the proposed framework integrates Explainable AI techniques such as Grad-CAM to visualize the regions of MRI images that contribute most significantly to classification decisions. Feature importance analysis methods are further applied to ensure transparency in model behaviour. This module assists clinicians in understanding model predictions and enhances trust in AI-assisted diagnosis.

Model Evaluation and Optimization: The performance of the proposed ensemble framework is assessed using multiple evaluation metrics including Accuracy, Precision, Recall, F1-

score, and ROC-AUC.Stratified k-fold cross-validation is implemented to ensure unbiased performance estimation, particularly in handling class imbalance. Hyperparameter tuning using Bayesian optimization is employed to enhance model performance and achieve optimal parameter configuration.

VI. RESULTS AND DISCUSSION

The proposed explainable ensemble transfer learning framework was evaluated using a stratified 5-fold cross-validation approach to ensure statistical robustness. Bayesian optimization was applied for hyperparameter tuning of individual models prior to ensemble integration. Experimental results demonstrate that the ensemble model consistently outperforms individual deep learning architectures across all evaluation metrics. The integration of CNN and Transformer-based models enables comprehensive spatial and contextual feature learning. The inclusion of Explainable AI further validates model decisions by highlighting tumour-relevant regions in MRI scans. Comparative analysis confirms improved classification performance across all four classes, demonstrating the effectiveness of the proposed architecture in handling multiclass brain tumour detection tasks.

VII. CONCLUSION AND FUTURE WORK

This study presents an explainable ensemble transfer learning framework for multiclass brain tumour classification using MRI images. By integrating pre-trained deep learning architectures with ensemble learning and explainability mechanisms, the system achieves high diagnostic accuracy and improved model transparency. The framework effectively addresses challenges such as limited medical data availability, class imbalance, and model interpretability. The results indicate that ensemble-based transfer learning significantly enhances classification robustness compared to standalone models. Future work may focus on incorporating larger multi-institutional MRI datasets to further improve generalization. Additionally, real-time deployment in clinical environments and integration with hospital information systems can be explored. Advanced explainability techniques and lightweight model architectures may also be investigated to support edge-based medical diagnostic systems.

REFERENCES

1. N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, “A deep learning model based on concatenation approach for the diagnosis of brain tumor,” IEEE Access, vol. 8, pp. 55135–55144, 2020.
2. A. Wulandari, R. Sigit, and M. M. Bachtiar, “Brain tumor segmentation to calculate percentage tumor using MRI,” in Proc. IEEE Int. Electron. Symp. Knowl. Creation Intell. Comput., 2018, pp. 292–296.
3. E. S. Chahal, A. Haritosh, A. Gupta, K. Gupta, and A. Sinha, “Deep learning model for brain tumor segmentation and analysis,” in Proc. IEEE 3rd Int. Conf. Recent Develop. Control, Autom. Power Eng., 2019, pp. 378–383.
4. C. Wang et al., “Phenotypic and genetic associations of quantitative magnetic susceptibility in U.K. biobank brain imaging,” Nature Neuroscience, vol. 562, pp. 1–14, May 2022.
5. M. I. Sharif, J. P. Li, M. A. Khan, and M. A. Saleem, “Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images,” Pattern Recognition Letters, vol. 129, pp. 181–189, Nov. 2019.
6. Z. Liu et al., “Deep learning based brain tumor segmentation: A survey,” Complex & Intelligent Systems, vol. 9, no. 1, pp. 1001–1026, 2020.
7. M. A. Ottom, H. A. Rahman, and I. D. Dinov, “ZNet: Deep learning approach for 2D MRI brain tumor segmentation,” IEEE J. Translational Engineering in Health and Medicine, vol. 10, Art. no. 1800508, May 2022.
8. H. S. Abdulbaqi, K. N. Mutter, M. Z. M. Jafri, and Z. A. Al-Khafaji, “Estimation of brain tumour volume using expanded computed tomography scan images,” in Proc. IEEE 23rd Iranian Conf. Biomed. Eng., 2016, pp. 117–121.
9. R. Sethi, M. Mehrotra, and D. Sethi, “Deep learning based diagnosis recommendation for COVID-19 using chest X-rays images,” in Proc. IEEE 2nd Int. Conf. Inventive Res. Comput. Appl., 2020, pp. 1–4.
10. S. V. Militante, N. V. Dionisio, and B. G. Sibbaluca, “Pneumonia detection through adaptive deep learning models of convolutional neural networks,” in Proc. IEEE 11th Control Syst. Graduate Res. Colloq., 2020, pp. 88–93.
11. S. Hussein, P. Kandel, C. Bolan, M. B. Wallace, and U. Bagci, “Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches,” IEEE Trans. Medical Imaging, vol. 38, no. 8, pp. 1777–1787, Aug. 2019.
12. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, “Chest pathology detection using deep learning with non-medical training,” in Proc. IEEE 12th Int. Symp. Biomed. Imaging, 2015, pp. 294–297.
13. Z. Lv, L. Qiao, and A. K. Singh, “Advanced machine learning on cognitive computing for human behavior analysis,” IEEE Trans. Computational Social Systems, vol. 8, no. 5, pp. 1194–1202, Oct. 2021.
14. T. M. Ali et al., “A sequential machine learning-cum-attention mechanism for effective segmentation of brain tumor,” Frontiers in Oncology, vol. 12, Jun. 2022.
15. M. Bhuvaneswari, “Automatic segmenting technique of brain tumors with convolutional neural networks in MRI images,” in Proc. IEEE 6th Int. Conf. Inventive Comput. Technol., 2021, pp. 759–764.

16. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. 25th Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
17. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. 3rd Int. Conf. Learn. Representations, Apr. 2015, pp. 1–14.
18. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, “Lung pattern classification for interstitial lung diseases using a deep convolutional neural network,” IEEE Trans. Medical Imaging, vol. 35, no. 5, pp. 1207–1216, Feb. 2016.
19. C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” IEEE Trans. Multimedia, vol. 17, pp. 2049–2058, 2015.
20. Q. Wang, F. Liu, G. Wan, and Y. Chen, “Inference of brain states under anesthesia with meta learning based deep learning models,” IEEE Trans. Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1081–1091, 2022.