JSL, 8(6), JUNE 2023: 16-24

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

JOURNAL OF SCIENTIFIC LETTERS
www.jslsci.com

MULTIMODAL DEEP LEARNING FOR AUTOMATED AND EXPLAINABLE CLASSIFICATION OF BENIGN AND MALIGNANT TUMORS USING INTEGRATED MEDICAL IMAGING DATA

Ninad N Thorat

Research Scholar, Sunrise University, Alwar, Rajastha

Dr Ashok Kumar

Professor, Sunrise University, Alwar, Rajastha

ABSTRACT

Conventional tumor assessment primarily depends on radiologists' interpretation of medical images, which may be affected by image quality, observer variability, and the complexity of tumor morphology. Recent advances in deep learning have created opportunities for automated and reliable tumor classification, particularly through convolutional neural networks and multimodal learning architectures. This paper presents a conceptual framework for multimodal deep learning that integrates complementary medical imaging information to distinguish benign from malignant tumors. The proposed approach combines image preprocessing, modality-specific feature extraction, multimodal feature fusion, classification, and explainability mechanisms. By integrating information from different imaging modalities, the framework can capture structural, textural, and functional characteristics that may not be sufficiently represented by a single modality. Explainable artificial intelligence techniques are incorporated to highlight diagnostically relevant image regions and provide interpretable predictions. The framework aims to improve classification accuracy, robustness, transparency, and clinical usefulness in computer-assisted tumor diagnosis.

Keywords: Multimodal Deep Learning, Tumor Classification, Benign Tumor, Malignant Tumor, Medical Imaging, Explainable AI, Convolutional Neural Networks, Image Fusion, Computer-Aided Diagnosis, Deep Learning

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

I. INTRODUCTION

Tumors represent a major clinical challenge because their biological behavior can range from relatively harmless benign growths to aggressive malignant neoplasms. Accurate identification of malignancy is fundamental to determining appropriate clinical management, selecting therapeutic strategies, and estimating prognosis. Medical imaging plays an important role in this process because it enables non-invasive visualization of tumor location, morphology, tissue characteristics, and structural abnormalities. Modalities such as magnetic resonance imaging (MRI), computed tomography (CT), mammography, ultrasound, positron emission tomography (PET), and histopathological imaging provide different forms of information about tumors. However, interpreting these images can be complex, particularly when lesions exhibit heterogeneous characteristics or overlap in appearance between benign and malignant conditions.

Traditional computer-aided diagnosis systems generally depend on manually engineered features, including shape, intensity, texture, edge characteristics, and statistical descriptors. Although such methods have contributed to automated image analysis, their performance may depend heavily on feature-selection strategies and domain-specific assumptions. Deep learning has substantially transformed medical image analysis by enabling computational models to learn hierarchical representations directly from imaging data. Convolutional neural networks (CNNs), in particular, have demonstrated strong performance in classification, segmentation, detection, and feature extraction tasks.

Despite these advances, single-modality systems have an inherent limitation because each imaging modality captures only particular aspects of tumor biology. MRI may provide excellent soft-tissue characterization, CT can reveal anatomical density and structural information, while PET can provide metabolic information. Consequently, combining complementary modalities can potentially generate a more comprehensive representation of a tumor. Multimodal deep learning addresses this challenge by learning features from multiple sources and integrating them through early, intermediate, or late fusion strategies.

Another important limitation of automated diagnostic systems is their lack of interpretability. A model may provide a highly accurate prediction while offering little information about why a particular lesion has been classified as benign or malignant. In clinical environments, such opacity can reduce physician confidence and complicate validation. Explainable artificial

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

intelligence (XAI) methods, including gradient-based visualization, saliency mapping, and attention mechanisms, can help identify image regions that influence predictions.

The present study therefore focuses on a multimodal deep learning framework for automated and explainable classification of benign and malignant tumors using integrated medical imaging data. The proposed framework emphasizes complementary information extraction, multimodal feature fusion, robust classification, and interpretable decision-making. Such an approach has the potential to support radiologists and clinicians by providing consistent, quantitative, and transparent computer-assisted diagnostic information.

II. MULTIMODAL DEEP LEARNING FRAMEWORK FOR TUMOR CLASSIFICATION

A multimodal deep learning framework for tumor classification can be designed as a sequential pipeline consisting of data acquisition, preprocessing, modality-specific feature extraction, multimodal feature fusion, classification, and interpretability analysis. The central objective is to transform heterogeneous medical imaging data into a unified representation capable of accurately distinguishing benign from malignant tumors. The effectiveness of the framework depends not only on the selected deep learning architecture but also on appropriate preprocessing, registration, normalization, data balancing, and fusion strategies.

The first stage involves collecting imaging datasets containing confirmed benign and malignant tumor cases. Depending on the clinical application, the dataset may include MRI, CT, ultrasound, mammography, PET, or combinations of these modalities. Each modality provides complementary information. For example, structural imaging can reveal lesion size and boundaries, whereas functional imaging may indicate metabolic activity. The availability of multiple modalities therefore enables the model to learn a richer representation of tumor characteristics. Reliable diagnostic labels should preferably be obtained from established clinical assessments, pathological confirmation, or expert consensus.Medical images frequently contain variations in spatial resolution, intensity distribution, orientation, acquisition protocol, and anatomical coverage. Therefore, preprocessing is necessary before model training. Common operations include image resizing, intensity normalization, noise reduction, contrast enhancement, artifact correction, and region-of-interest extraction. When images from different modalities correspond to the same patient, image registration can be applied to establish spatial correspondence. Data augmentation techniques such as rotation,

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

translation, flipping, scaling, cropping, and controlled intensity transformation can further improve generalization and reduce over fitting.

Following preprocessing, modality-specific deep learning branches can be employed. CNN-based architectures are particularly appropriate because they can automatically learn spatial and hierarchical representations from medical images. More recent architectures, including vision transformers and hybrid CNN-transformer networks, may also be incorporated to capture both local and long-range contextual information. Each branch processes its corresponding imaging modality and generates a latent feature representation. Modality-specific learning is advantageous because different imaging sources have distinct statistical properties and clinically meaningful patterns.The extracted representations are subsequently combined through multimodal fusion. Three major strategies can be considered. Early fusion combines raw or minimally processed modality data before feature extraction. Although relatively simple, this approach may be sensitive to differences in resolution and modality characteristics. Intermediate fusion extracts modality-specific features first and then combines them at a deeper representation level. This strategy can preserve modality-specific information while enabling cross-modal interactions. Late fusion independently generates predictions from each modality and combines the outputs using weighted averaging, voting, or another decision-level mechanism. Attention-based fusion can further enable the network to assign greater importance to informative modalities or image regions.

The fused representation is then provided to fully connected or classification layers that estimate the probability of benign and malignant categories. A binary classification function can produce a malignancy probability, while threshold optimization can be performed according to clinical objectives. Evaluation should include accuracy, precision, recall, specificity, F1-score, sensitivity, area under the receiver operating characteristic curve (AUC-ROC), and confusion matrices. Sensitivity is particularly important because failure to identify a malignant tumor may have serious clinical consequences. Model validation should employ patient-level training, validation, and testing partitions to prevent data leakage. Cross-validation can provide additional estimates of model stability, while external validation on an independent dataset is highly desirable for assessing generalizability. Class imbalance can be addressed through balanced sampling, augmentation, weighted loss functions, or focal loss.

Overall, multimodal deep learning provides a flexible computational framework capable of integrating complementary imaging evidence. Rather than relying on one source of visual

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

information, the framework learns relationships among different representations and potentially improves the reliability of automated tumor classification. However, its success depends on high-quality datasets, appropriate fusion mechanisms, careful validation, and clinically meaningful interpretability.

III. EXPLAINABLE ARTIFICIAL INTELLIGENCE AND CLINICAL INTERPRETATION OF MULTIMODAL TUMOR PREDICTIONS

Although classification accuracy is an important measure of a medical artificial intelligence system, accuracy alone is insufficient for clinical adoption. Physicians need to understand the basis of an automated prediction, particularly when the model identifies a lesion as malignant. Explainable artificial intelligence provides mechanisms for interpreting model behavior and identifying the visual evidence associated with predictions. In the proposed multimodal tumor-classification framework, explain ability should therefore be incorporated as an integral component rather than treated as an optional post-processing procedure.

One of the most widely used approaches for interpreting deep learning models is class activation mapping. Methods such as Grad-CAM generate visual heatmaps that indicate image regions contributing strongly to a particular prediction. When applied to tumor classification, these maps can demonstrate whether the model is focusing on the lesion itself, its boundaries, internal heterogeneity, surrounding tissue, or irrelevant artifacts. Such information can help clinicians assess whether the model has learned medically meaningful features. If the prediction is based primarily on irrelevant regions, the visualization can reveal potential weaknesses in the model.

Saliency-based techniques provide another mechanism for identifying influential image features. They estimate how changes in image pixels or representations affect the model output. Integrated gradients and related attribution approaches can provide additional information about the contribution of specific regions to classification. In multimodal systems, explainability can be extended across modalities to determine which imaging source contributes most strongly to a prediction. This is particularly valuable when modalities contain complementary or occasionally conflicting information.

Attention mechanisms provide a further opportunity for interpretability. An attention-based multimodal architecture can learn to emphasize informative spatial regions and modalities. For example, if a particular imaging modality provides stronger evidence of malignancy, the

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

model may assign greater weight to its representation. Attention maps can then be examined to understand the spatial and modality-level distribution of model focus. However, attention weights should not automatically be interpreted as definitive causal explanations; they should be considered complementary evidence alongside other interpretability techniques.A clinically useful explainable system should also provide confidence estimates. Instead of reporting only "benign" or "malignant," the model can generate a probability score accompanied by visual explanations. A high-confidence malignant prediction supported by a localized lesion-focused heatmap may provide more useful information than an unexplained categorical output. Nevertheless, model confidence should not be equated with clinical certainty because neural networks can be confidently incorrect when exposed to unfamiliar data.

Explainability is also important for detecting dataset bias and shortcut learning. Deep learning models may unintentionally learn characteristics associated with imaging equipment, acquisition protocols, hospitals, demographic distributions, or image annotations instead of tumor biology. Visualization methods can help researchers identify such behavior. For example, if a heatmap consistently highlights scanner artifacts rather than tumor regions, the model may be relying on non-clinical correlations. Therefore, explainability can support model auditing and quality assurance.The proposed framework can incorporate multiple levels of interpretation. At the local level, explanations describe why an individual tumor has been classified as benign or malignant. At the global level, researchers can analyze patterns of model behavior across the dataset to determine whether the network consistently identifies clinically relevant characteristics. Feature-importance analysis can also be used to compare the contributions of different imaging modalities.

Clinical evaluation should involve radiologists, oncologists, or other relevant specialists. Experts can assess whether highlighted regions correspond to recognizable diagnostic characteristics and whether model predictions provide useful supplementary information. Human-AI comparison and collaboration studies may further determine whether the system improves diagnostic consistency or reduces interpretation time.Despite its benefits, explainability has limitations. Different XAI techniques can generate different explanations for the same prediction, and a visual heatmap does not necessarily establish a causal relationship between a region and the classification decision. Consequently, explanations should be validated rather than accepted uncritically. Furthermore, explanations must be

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

presented in a manner understandable to clinicians without creating false confidence.

The integration of explainability with multimodal learning can therefore contribute to a more transparent computer-aided diagnostic system. Such a system does not seek to replace clinical expertise but rather to provide an additional source of quantitative evidence. By combining multimodal imaging, automated classification, and interpretable visualization, the proposed framework can support more transparent and potentially more reliable tumor assessment.

IV. CONCLUSION

Multimodal deep learning represents a promising approach for improving automated differentiation between benign and malignant tumors. Medical imaging modalities contain complementary information regarding anatomical structure, tissue characteristics, lesion morphology, and, in some cases, physiological or metabolic behavior. Conventional single-modality approaches may fail to capture the full complexity of these characteristics, whereas multimodal architectures can integrate heterogeneous information into a unified representation. The proposed framework combines preprocessing, modality-specific deep feature extraction, multimodal fusion, classification, and explainability to establish a comprehensive approach to automated tumor assessment.

The use of deep learning enables the system to learn complex imaging patterns without depending exclusively on manually engineered features. Modality-specific branches can preserve the unique characteristics of each imaging source, while intermediate or attention-based fusion can identify relationships between complementary representations. Appropriate evaluation using sensitivity, specificity, precision, recall, F1-score, and AUC can provide a comprehensive assessment of diagnostic performance. Patient-level data separation and independent external validation are particularly important to ensure that reported performance reflects genuine generalization rather than data leakage or overfitting.

Explainability is equally important because clinical decision-making requires transparency and accountability. Techniques such as Grad-CAM, saliency analysis, integrated gradients, and attention visualization can identify image regions and modalities associated with model predictions. These mechanisms may help clinicians understand automated outputs, identify potentially incorrect model behavior, and assess whether the system focuses on clinically meaningful tumor characteristics. Nevertheless, explainable AI should complement rather than replace expert judgment.

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

Future development should focus on larger and more diverse datasets, standardized multimodal acquisition protocols, external validation across institutions, uncertainty estimation, privacy-preserving learning, and prospective clinical evaluation. Federated learning may provide opportunities to train models across institutions without directly sharing sensitive patient data. Furthermore, combining imaging information with clinical, genomic, pathological, or laboratory data could create more comprehensive multimodal diagnostic systems.

In conclusion, the integration of multimodal deep learning and explainable artificial intelligence offers a strong foundation for developing reliable computer-assisted tumor classification systems. By combining diverse medical imaging evidence with transparent predictive mechanisms, such systems have the potential to enhance diagnostic support, improve consistency, and facilitate more informed clinical decision-making. However, successful clinical translation will require rigorous validation, interpretability assessment, ethical oversight, and continued collaboration between artificial intelligence researchers and medical professionals.

V. REFERENCES

1. G. Litjens et al., “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

2. A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017.

3. G. J. S. Litjens et al., “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis,” *Scientific Reports*, vol. 6, Art. no. 26286, 2016.

4. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

5. 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 (CVPR)*, 2016, pp. 770–778.

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

6. R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 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. S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

9. M. H. Yap et al., “Automated breast ultrasound lesions detection using convolutional neural networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1215–1224, 2018.

10. J. P. Kaushal, A. K. K. Bhatia, and related literature on multimodal medical image analysis, emphasizing the integration of complementary imaging information for computer-aided diagnosis.