National Conference on Behavioural Economics and Intelligent Decision Systems for Climate Change and Sustainable Development
Journal of Advanced Research in Medical Science & Technology
Volume 12, Issue 3&4 - 2025

Review Article

A review of Multimodal-based Deep Learning Architectures

Sunaina[1], Baljit Kaur[2], Priya Thakur[3], Navreet Kaur[4]

[1,2,3,4] Research Scholar, Department of Computer Science and Engineering , DAV University, Jalandhar, Punjab, India
DOI: https://doi.org/10.24321/2394.5439.202506

I N F O

Corresponding Author:
Sunaina, Department of Computer Science and Engineering , DAV University, Jalandhar, Punjab, India
E-mail Id:
sunaina3086@gmail.com
How to cite this article:
Sunaina, Kaur B, Thakur P, Kaur N. A review of Multimodal-based Deep Learning Architectures. J Adv Res Med Sci Tech. 2025;12(3&4):222-228.

Date of Submission: 2025-10-04
Date of Acceptance: 2025-10-29

A B S T R A C T

Multimodal deep learning has emerged as a significant approach in medical imaging. It allows the incorporation of complementary data derived from multiple imaging modalities, such as CT, MRI, and PET. This review looks at recent developments in deep learning structures that combine multiple modalities, like CT with MRI, PET with MRI, and CT with PET, to improve disease diagnosis and prognosis. These fusion methods capture both anatomical and functional details. As a result, models can learn richer feature representations that lead to better accuracy and reliability. Structures such as convolutional neural networks, attention-based networks, generative adversarial networks (GANs), and hybrid fusion frameworks have performed exceptionally well in tasks like tumour segmentation, disease classification, and mutation prediction. Studies show notable improvements in diagnosing complex conditions, including lung cancer, brain tumours, Alzheimer's disease, and esophageal cancer. Additionally, integrating explainable AI methods increases transparency and clarity in clinical decisions. Overall, this review highlights that multimodal deep learning, using effective fusion of techniques like CT and MRI or PET and MRI, is advancing toward more precise, timely, and personalised medical diagnosis.

Keywords: Multimodal Deep Learning, Medical Image Fusion, CT–MRI–PET Integration, Disease Diagnosis, Attention Mechanisms, Explainable AI

Introduction

Medical imaging tools like CT for bone details, MRI for soft tissues, and PET for body activity help doctors detect diseases such as lung cancer, brain tumours, and Alzheimer's disease at an early stage. Each modality has its own strengths and limitations. CT provides structural detail but lacks functional information, while MRI captures soft tissue contrast but omits bone structures, and PET offers functional data with limited spatial precision. By integrating these complementary imaging modalities, clinicians can obtain a comprehensive understanding of both structural and functional aspects of disease pathology.[1,2,3]

Deep learning AI makes this fusion easy with smart networks like CNNs and GANs that pull key features from mixed scans. For lung cancer, it predicts gene changes from PET/CT with 88% accuracy;[6] for brain tumours, it sorts types from CT/MRI at 99% right.[2] In Alzheimer's, PET-MRI blends spot brain shrinkage and low energy use at 74% success, using tools to explain why.[4,8] These speed up analysis and cut mistakes.

Key models include

- **MedClip/BEiT Fusion Model:**[1] Merges CT/PET scans with clinical/genomic info to classify NSCLC using MedClip for vision-language features and BEiT for pre-

Copyright (c) 2025: Author(s). Published by Advanced Research Publications

training; denoising is done using a CNN autoencoder and wavelet decomposition.

- **MBTC-Net:**[2] Brain tumour classification from CT and MRI (T1/T1CE/T2) is done with an EfficientNetV2B0-based architecture network with multi-head attention focused on layer context; spectrum reshaping and 5-fold validation were fine-tuned using Adamax with Grad-CAM explainability.
- **Attention-Based Inception-ResNet:**[3] Lung cancer detection across X-ray/CT/histopathology is done using a hybrid; Inception for multi-scale features, ResNet residuals, and channel attention focus via multi-modal fusion.
- **Heuristic ResNet18 Fusion Model:**[4] For binary AD classification, early fusion of PET/MRI via 3-channel concatenation; modified ResNet18 manages heterogeneity with XAI (Grad-CAM) while correlating atrophy and metabolism.
- **MedFusionGAN:**[5] Unsupervised GAN for fusion of CT (bones) and 3D T1-Gd MRI (soft tissues) of brain tumours; The generator uses content/style/L1 losses on the GLIS-RT dataset for radiotherapy delineation.
- **3D CNN Transfer Learning Model (TS_TL):**[6] The PET/CT and clinical data to predict EGFR mutation in lung adenocarcinoma incorporate a 3D CNN with transfer learning in a 3-stream architecture where segmentation/resampling is included.

These architectures improve some metrics—precision, recall, or F1-scores—along with explainability, which was emphasised as necessary for clinical use.[1,4] Conducting imaging, radiomics, and clinical data, Deep Learning (DL) multimodal systems, such as DL-Radiomics-Clinical models for ESCC metastasis prediction, prove external validity with an AUC of 0.916.[7] Although there are still some data heterogeneity, class imbalance, and computational workload, improvements in unsupervised learning and transfer of learning AI techniques seem to resolve these issues.[5,6] For these reasons, deep learning multimodal architecture design is a key shift for accurate and speedy diagnosis, and, as a result, diagnosis tools are provided to the clinicians for timely intervention and enhanced patient treatment in oncology or neurology.

The figure represents a general deep learning architecture employed for multimodal medical image fusion. It consists of data preprocessing, feature extraction, multimodal fusion, and classification stages. As shown in Figure 1, deep learning architectures combine data from different imaging modalities like CT, MRI, and PET to extract anatomical as well as functional features for enhanced detection of diseases.

Overall, these methods help doctors outline tumours better, track treatment changes, and spot spread early, like in throat cancer nodes.[7,12]

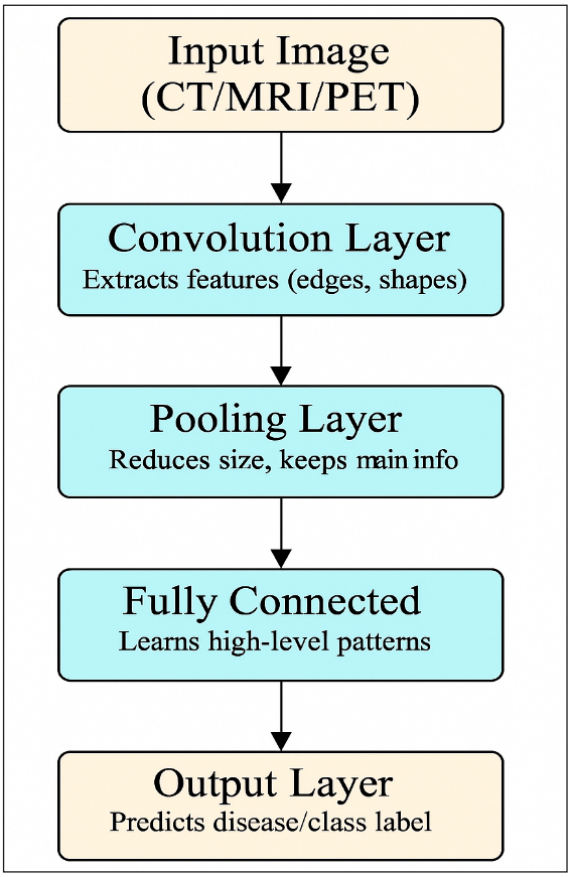

Figure 1.Deep Learning Architecture

Literature Survey

Recent research in medical imaging has focused on multimodal deep learning. This approach combines information from different imaging methods like CT, MRI, and PET. By integrating these methods, it captures both structural and functional features, which helps improve disease detection and classification accuracy. Studies show that fusion-based deep learning models perform better than single-modality approaches. They provide more reliable and complete diagnostic insights for various diseases. The following literature review summarises and analyses key contributions from 2022 to 2025 that exemplify multimodal-based deep learning architectures:

Hassan S et al.[1] A multi-modal fusion approach for non-small cell lung cancer (NSCLC) classification integrates CT and PET scans with patient-related and genetic information. The technique uses MedClip and BEiT models to extract image features from denoised scans with a deep CNN auto-encoder. It also involves wavelet decomposition and image registration for fusion, alongside preprocessing of tabular and genetic data. This preprocessing includes handling missing values, encoding, balancing, scaling, and selecting features. A multi-modal classifier then combines these features for detection and subtype classification. An advantage is its accuracy of 94.04% across metrics like

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506

precision, recall, and F1-score, which is better than single-modality approaches, as it captures both anatomical and metabolic tumour aspects. A disadvantage is the heavy reliance on high-quality genomic data, which may limit its use in resource-limited settings without thorough patient records.

Similarly, Kar S et al.[2] MBTC-Net classifies brain tumours based on CT and MRI (T1, T1CE, T2) scans. It utilises EfficientNetV2B0 with multi-head attention to extract contextual features, followed by pooling, normalisation, and dense layers optimised through Adamax and softmax. With fivefold cross-validation, it achieves a maximum of 99.34% accuracy for Kaggle datasets, although its multi-head attention mechanism places an additional computational burden on non-GPU environments. An advantage is its high accuracies (up to 99.34% for binary classification) across diverse Kaggle datasets, allowing for strong generalisation in real-time use. A disadvantage is the computational intensity of the multi-head attention mechanism, which can slow efficiency on standard hardware without GPU acceleration.

Hosny M et al.[3] A multi-modal deep learning framework for lung cancer detection combines X-rays, CT, and histopathological images. The approach uses an Inception-ResNet module for multi-scale spatial feature extraction with residual learning for optimising gradients, enhanced by the model integrating multi-level convolutional fusion in sequence with an optimised channel attention mechanism that prioritises important regions. Classification is done on large datasets. An advantage is its accuracy (up to 99.73% on CT data), surpassing traditional deep learning models and improving clinical practice through attention-focused feature enhancement. A disadvantage is the need for large, balanced datasets for training, which could reduce performance on smaller or imbalanced real-world groups.

Odusami M et al.[4] An explainable deep-learning framework diagnoses Alzheimer's disease (AD) by fusing PET and MRI images. It involves early feature fusion by concatenating PET data, which measures glucose metabolism, and MRI data, which assesses brain volume atrophy, into a three-channel input. This is trained on a modified ResNet18 architecture to learn features for binary AD classification, using Explainable AI for result interpretation. An advantage is its handling of data diversity, achieving 73.90% accuracy on the ADNI database and providing interpretable insights into model decisions. A disadvantage is its relatively modest accuracy compared to unimodal methods in some cases, affecting its reliability for standalone diagnosis. metrics. A disadvantage is the potential for inconsistent intensity profiles in complex tumour areas, which may require manual adjustments for precise delineation.

Safari M et al.[5] MedFusionGAN is for fusing in brain tumor tumourradiotherapy planning. This method consists of a generator and discriminator jointly trained with content, style, and L1 losses to preserve\tstructure. tumourpreserve structure. employs a multicenterpreserve structure. multicentre dataset (GLIS-RT) including different protocols and tumourmulticentrer types for generalisation. generalisation. The benefit is its unsupervised character, generating coalesced images with higher spatial resolution and fewer artefacts, outperforming seven conventional and eight deep learning techniques based on six of nine metrics.

A drawback is that it can result in variable intensity profiles in complicated tumour regions, which will need to be manually corrected for accurate delineation.

Shao X et al.[6] A transfer learning-based 3D CNN predicts EGFR mutations in lung adenocarcinoma using PET/CT images merged with clinical data. This technique develops dual-stream (PET/CT) and three-stream models that integrate semiautomatic segmentation, resampling, cropping, and class imbalance adjustments. It is fine-tuned on data from 516 patients, with ROC-AUC evaluation identifying activation patterns in lesions changed by tyrosine kinase inhibitors. An advantage is its clinical usefulness, with AUCs reaching 0.883 in training and 0.760 in advanced cases, offering insights for personalised therapy. A disadvantage is the limitation of using retrospective datasets, which may introduce selection bias affecting external validation.

Yuan P et al.[7] A hybrid deep learning–radiomics–clinical model was proposed to predict cervical lymph node metastasis of oesophageal squamous cell carcinoma based on ^{18}F-FDG PET/CT images. The system extracts 428 quantitative radiomic features from the original tumour area and cervical lymph nodes layer-by-layer, along with 256 deep learning features from maximum cross-sections. These features are combined with clinical data across seven machine learning algorithms, with logistic regression selected as the classifier, tested on internal and external cohorts of 411 patients. An advantage is its high predictive AUCs (0.955 internal, 0.916 external), allowing for better non-invasive staging than single-modality models. A disadvantage is the complexity of extracting features from layered volumes, which increases preprocessing time and computational demands.

Song J et al.[8] A multimodal image fusion method for AD diagnosis uses MRI and FDG-PET. This technique registers and masks grey matter tissue from MRI onto PET images to create a "GM-PET" fused modality that emphasises crucial AD areas. It is evaluated with a 3D Simple Convolutional Neural Network and a 3D Multi-Scale CNN for binary and multi-class tasks on the ADNI dataset. An advantage is its interpretability and improved performance over unimodal

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506

and feature fusion methods, retaining both contour and metabolic details for accurate classification. A disadvantage is the reliance on precise registration, which can lead to alignment errors in diverse brain anatomies.

Chen W et al.[9] Res2Net-based MRI and CT image fusion is used for delineating intracranial tumours. This technique utilises Res2Net as a feature extractor for multiscale features and a spatial mean attention fusion layer that adaptively weighs to maintain details. It includes a reconstructor for the final image, evaluated on qualitative and quantitative metrics like average gradient and entropy. An advantage is that it preserves texture and structural information, exceeding advanced algorithms in visual quality and reducing spectral artefacts for clinical use. A disadvantage is the requirement for high-resolution source images, which could compromise fusion quality when working with lower-quality scans common in routine practice.

Alshmrani GM et al.[10] Hyper Dense Lung Seg is a modified U-Net for lung tumor segmentation using CT-PET multimodality. This method uses early, late, and dense, strategies, with being the best. It integrates functional and anatomical information from PET/CT scans from different datasets automatically. A benefit is its Dice score of 73%, used for early detection and performing better than other fusion techniques for clinicians. A drawback is the challenge of handling spectral and temporal differences between modalities, which can lead to inconsistencies in segmenting heterogeneous tumours.

Bhutto et al.[11] suggested a CT and MRI medical image fusion method that combines a noise-removal and contrast enhancement mechanism with a Convolutional Neural Network (CNN). The approach improves image resolution by suppressing artefacts and enhancing texture and edge features, facilitating enhanced diagnostic visualisation. It efficiently fuses the structural information of CT and the soft-tissue features of MRI. An advantage of this paper is that it offers high-quality fused images with enhanced contrast and minimised noise, and a disadvantage is Consumes high computational resources for training and processing of CNN.[12-15]

Rahul Hans et al. In this paper,a hybrid optimisation method referred to as SCALO, which is an integration of Sine Cosine Algorithm (SCA) and Ant Lion Optimiser (ALO), is proposed to perform feature selection. The approach promotes an enhanced exploration and exploitation trade-off, enhancing the search for the best feature subsets and enhancing classification precision. An advantage of this model is it efficiently prevents local minima and enhances convergence rate, and a disadvantage is the hybrid approach increases computational complexity due to its iterative nature.

In summary, the literature reviewed illustrates that multimodal deep learning models, which combine CT, MRI, and PET scans, perform much better than unimodal approaches in disease diagnosis.

A key comparison of such methods illustrates that early fusion methods tend to perform better when modalities are spatially well-aligned since they enable the model to learn joint feature representations at the input stage. Nonetheless, late fusion methods have more freedom to fuse heterogeneous data sources, e.g., clinical and genomic attributes, at the decision level. Hybrid fusion methods compromise between the two methods, with performance optimally balanced across modalities. Nevertheless, class imbalance, dataset heterogeneity, and the absence of cross-site validation remain real-world limitations. Largely resolving these problems needs bigger, standard multimodal datasets and solid validation practices across several institutions to provide clinical reliability.

Conclusion and Future Directions

Multimodal deep learning has shown great potential in improving disease diagnosis by combining complementary information from different imaging modalities such as CT, MRI, and PET. This integration helps capture both anatomical and functional details, leading to more accurate detection, segmentation, and classification of diseases. Compared to single-modality models, multimodal architectures reduce diagnostic uncertainty and enhance model generalisation across diverse clinical scenarios. However, challenges such as image alignment, data imbalance, and computational complexity still limit their widespread clinical use.

Future research should focus on developing rotationally invariant multimodal architectures that can handle orientation variations across medical images. Integrating rotational invariance can make deep learning models more robust to image misalignment and scanner orientation differences, improving fusion accuracy. Expanding multimodal datasets and incorporating explainable AI techniques will further enhance model reliability and clinical interpretability, paving the way for real-time and generalisable diagnostic systems.

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506

Table 1.Comparative Analysis of Multimodal Deep Learning

S.No.	Title	Imaging Modalities Used	Deep Learning Architecture / Model Used	Performance Metrics	Limitations / Challenges
1	MULTI-MODAL MEDICAL IMAGE FUSION FOR NON-SMALL CELL LUNG CANCER	CT + PET	BEiT-based multimodal classifier (with MedCLIP)	Accuracy: 94.04%; Precision, Recall, F1-Score	High computational requirements, limited cross-clinic scalability, data privacy concerns
2	MBTC-Net: Multimodal brain tumor classification from CT and MRI scans using deep neural network and attention mechanism	CT + MRI	MBTC-Net (EfficientNetV2B0 with multi-head attention)	Accuracy: 97.54%; F1: 97.53%; Precision: 97.53%; Recall: 97.54%	Noisy/irregular features due to inter-modality alignment error; overfitting for small datasets
3	Multi-Modal Deep Learning for Lung Cancer Detection Using Attention-Based Inception-ResNet	CT + X-ray / Histopathological Images	Attention-Based Inception-ResNet with Hierarchical Feature Extraction (Channel Attention)	Accuracy: 98.63%; Precision: 98.61%; Recall: 98.64%; F1: 98.63%	Higher computational requirements, data imbalance, small annotated datasets
4	Explainable Deep Learning-Based Diagnosis of Alzheimer's Disease Using Multimodal Fusion of PET and MRI Images	PET + MRI	In-3-Channel ResNet18 (Heuristic Early Feature Fusion Framework)	Binary Accuracy: 78.73%; Multi-class: 70.31%; F1: 73.65%	High computational requirements, data dependency, interpretability
5	MedFusionGAN: Multimodal medical image fusion using an unsupervised generative adversarial transfer learning-based framework	PET + CT	MedFusionGAN (Unsupervised GAN with ResNet-inspired generator and PatchGAN discriminator)	N/A (Fusion quality: SSIM, gradient loss, perceptual loss, generalization effectiveness 0.69)	Requires GPU training (116 min/sample), Limited dataset size, Complex architecture
6	PET/CT 3D convolutional neural network fusion of imaging and clinical information for prediction of EGFR mutation in lung cancer	PET + CT	TS_TL (Three-stream 3D CNN with transfer learning)	AUC: 0.739 (test), Sensitivity: 0.76, Specificity: 0.60	Requires diverse dataset, limited sample size, no external validation

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506

7	18F-FDG PET/CT-based deep learning radiomics-clinical model for prediction of cervical lymph node metastasis in esophageal squamous cell carcinoma	PET + CT	DRC model (ResNet50DL + Radiomics + Clinical features with Logistic Regression)	AUC: 0.955 (internal), 0.916 (external); Sensitivity: 0.951 (internal)	Relies on retrospective data, requires large dataset for generalization, imbalance in ROI extraction
8	An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis	MRI + PET	3D Simple CNN / 3D Multi-Scale CNN	Binary Accuracy: 84.14% (AD/NC), 84.83% (AD/MCI), 80.02% (MCI/NC)	Pre-processing intensive, needs balanced datasets
9	MR–CT image fusion method of intracranial tumors based on Res2Net	MRI + CT	Res2Net-based fusion (feature extraction, fusion, restoration with skip connection)	Average Gradient: 4.85; Entropy: 7.28; Spatial frequency: 16.37	Focused only on tumor regions; lacks multi-class validation
10	HyperDense_Lung_Seg: Multimodal-Fusion-Based Model for Lung Segmentation Using Multimodality of CT–PET	CT + PET	HyperDense network (VGG19 + U-Net with multimodal fusion approaches)	Dice: 78%; IoU: 68%; Sensitivity; AUC	Adjustment for PET and CT alignment needed; intensity normalization errors

References

1. Hassan S, Al Hammadi H, Mohammed I, Khan MH. Multi-modal medical image fusion for non-small cell lung cancer classification. arXiv preprint arXiv:2409.18715. 2024.
2. Kar S, Singh PK. MBTC-Net: Multimodal brain tumor classification from CT and MRI scans using deep neural network with multi-head attention mechanism. Medicine in Novel Technology and Devices. 2025;27:100382.
3. Hosny M, Elgendy IA, Albashrawi MA. Multi-Modal Deep Learning for Lung Cancer Detection Using Attention-Based Inception-ResNet. IEEE Access. 2025;13:123630–123651.
4. Odusami M, Maskeliūnas R, Damaševičius R, Misra S. Explainable Deep-Learning-Based Diagnosis of Alzheimer's Disease Using Multimodal Input Fusion of PET and MRI Images. Journal of Medical and Biological Engineering. 2023;43:291–302.
5. Safari M, Fatemi A, Archambault L. MedFusionGAN: multimodal medical image fusion using an unsupervised deep generative adversarial network. BMC Medical Imaging. 2023;23:203.
6. Shao X, Ge X, Gao J, Niu R, Shi Y, Shao X, Jiang Z, Li R, Wang Y. Transfer learning–based PET/CT three-dimensional convolutional neural network fusion of image and clinical information for prediction of EGFR mutation in lung adenocarcinoma. BMC Medical Imaging. 2024;24:54.
7. Yuan P, Huang ZH, Yang YH, Bao FC, Sun K, Chao FF, Liu TT, Zhang JJ, Xu JM, Li XN, Li F, Ma T, Li H, Li ZH, Zhang SF, Hu J, Qi Y. A 18F-FDG PET/CT-based deep learning-radiomics-clinical model for prediction of cervical lymph node metastasis in esophageal squamous cell carcinoma. Cancer Imaging. 2024;24:153.
8. Song J, Zheng J, Li P, Lu X, Zhu G, Shen P. An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. Frontiers in Digital Health. 2021;3:637386.
9. Chen W, Li Q, Zhang H, Sun K, Sun W, Jiao Z, Ni X. MR–CT image fusion method of intracranial tumors based on Res2Net. BMC Medical Imaging. 2024;24:169.

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506

10. Alshmrani GM, Ni Q, Jiang R, Muhammed N. Hyper-Dense_Lung_Seg: Multimodal-Fusion-Based Modified U-Net for Lung Tumour Segmentation Using Multimodality of CT-PET Scans. Diagnostics. 2023;13:3481.
11. Bhutto J. A.; Tian L.; Du Q.; Sun Z.; Yu L.; Tahir M. F., CT and MRI Medical Image Fusion Using Noise-Removal and Contrast Enhancement Scheme with Convolutional Neural Network, 2022;24(3):393.
12. Mao Q.; Zhai W.; Lei X.; Wang Z.; Liang Y., CT and MRI Image Fusion via Coupled Feature-Learning GAN, 2024;13(17):3491.
13. Chen W.; Li Q.; Zhang H.; et al., MR–CT image fusion method of intracranial tumors based on Res2Net, 2024;24:169.
14. Shunchao Guo; Lihui Wang; Qijian Chen; Li Wang; Jian Zhang; Yuemin Zhu, Multimodal MRI Image Decision Fusion-Based Network for Glioma Classification, 2022;12:819673.
15. Zhai et al. [5] “CT and MRI Image Fusion via Dual-Branch GAN (DBGAN),” 2023
16. Hans, R., & Kaur, H. (2020). Hybrid binary Sine Cosine Algorithm and Ant Lion Optimization (SCALO) approaches for feature selection problem. International Journal of Computational Materials Science and Engineering, 9(01), 1950021.

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.5439.202506