Multimodal Medical Image Fusion for Tumor Detection : A Survey

Kancharagunta Kishan Babu
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
kishan.kancharagunta@gmail.com

Nakka Smitha
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
nakkasmitha@gmail.com

Kore Abhijith
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
abhijithkore09@gmail.com

Palliyana Shabarish
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
palliyanashabarish004@gmail.com

Preetam Pujari
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
preetam.naik3@gmail.com

Balini Lohith Reddy
Department of CSE(AIML&IoT)
VNR VJIET
Hyderabad-500090, Telangana, India
lohithreddybalini@gmail.com

***Abstract*—Medical Image fusion (MIF) in has become a technique which integrates complementary information from multiple modalities to produce an enhanced composite image that helps with improved diagnostic capabilities and visual quality. In the context of detecting brain tumors in patients, commonly used modalities such as MRI, CT, PET, SPECT, fMRI, and DTI provide structural, functional, and metabolic insights that, when fused, offer a more comprehensive understanding of tumor characteristics. This survey presents an review of multi-image fusion techniques, discussing briefly the traditional methods such as weighted averaging, principal component analysis (PCA), and wavelet transforms and providing explanation on advanced deep learning-based models like Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Attention mechanisms. The paper highlights the principles and limitations of these approaches. The analysis reveals that while classical methods remain useful for certain applications, deep learning-based approaches show superior performance in handling complex multimodal data and various imaging conditions. Finally, the paper identifies existing research gaps and outlines promising future directions to guide advancements in MIF for brain tumor detection.**

***Index Terms*—Medical Image Fusion (MIF), Generative Adversarial Networks (GANs), Deep Learning (DL), Convolutional Neural Networks (CNNs), wavelet transforms, Attention mechanisms**

I. Introduction

A brain tumor is a medical condition where cells inside the brain or its surrounding structures grow uncontrollably. Both benign (non-cancerous) and malignant (cancerous) tumors have the potential to cause serious neurological impairments because they put pressure on critical brain regions. Higher survival rates, better patient prognoses, and efficient treatment planning are all facilitated by early tumor detection. Delays in detection cause irreversible brain tissue damage and make it more difficult to remove those areas surgically. This lessens the treatment's efficacy. As a result, accurate brain tumor localization and segmentation are essential for clinical decision-making.

Brain tumor diagnosis and monitoring depend heavily on medical imaging. Different image modalities capture different aspects of the brain. Tumor boundaries and edema can be seen with Magnetic Resonance Imaging (MRI) [1], which provides detailed structural information. In addition to MRI scans, computed tomography (CT) can be used to detect calcifications and bone involvement. Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) provide metabolic and functional information [2], aiding in the differentiation of benign from malignant tumors. Additional information on brain activity and white matter integrity is provided by Diffusion Tensor Imaging (DTI) and Functional MRI (fMRI), both of which are essential for surgical planning. Clinicians can gain a more comprehensive understanding of tumor behavior by integrating data from these modalities.

Combining complementary details from various modalities into a single, more informative image is known as image fusion. By combining the spatial resolution of one modality with the functional information of another, fusion aims to improve the interpretability and utility of medical data. While modern approaches use DL models to automate the learning of optimal fusion strategies, traditional fusion techniques include pixel-level, feature-level, and decision-level methods. The quality and comprehensiveness of diagnostic imaging are greatly enhanced by this integration.

Fused images help overcome the limitations of individual modalities in the analysis of brain tumors. PET shows functional activity, whereas MRI offers exact structural details but lacks metabolic information. It is possible to simultaneously visualize anatomical and functional information by combining MRI and PET images. Tumor localization, segmentation accuracy, and classification performance all improve as a result. Image fusion is a crucial method in contemporary neuroimaging and clinical practice since it helps with treatment planning, tracking therapy response, and lowering diagnostic uncertainty.

The paper presents a comprehensive survey of techniques

for fusing images for tumor detection in brain. It analyzes traditional fusion methods alongside recent advances in deep learning-based approaches and compares their strengths and limitations. Furthermore, the paper discusses the metrics for fusion quality assessment, metrics for tumor detection, and popular datasets used in the studies . By analyzing current research trends and identifying gaps, this study aims to guide future developments in medical image fusion for efficient brain tumor diagnosis.

II. Literature Review

This study explores many techniques used for the brain image fusion. In our study on fusion techniques for brain images, we looked into several traditional techniques which used transformations for fusion, and also deep learning approaches.

A. Traditional Techniques

Changtao He et al. [3] explored combining the IHS (Intensity–Hue–Saturation) color-space transform with Principal Component Analysis (PCA) [4] to fuse multimodal medical images such as MRI and PET; their pipeline separates color (functional information) from intensity, uses histogram matching and PCA to merge intensity components so MRI structural detail is preserved while PET contributes functional color cues, and then recombines channels back to RGB. This transform-plus-statistics approach directly addresses the common trade-off in early fusion methods between preserving spatial detail and retaining functional color information; by using IHS to isolate color and PCA to prioritise dominant intensity components it reduced color distortion and tried to keep anatomical detail. The limitation, however, remains that these hand-crafted rules cannot learn complex nonlinear relationships between modalities and can introduce artifacts when the modalities differ greatly in noise, resolution, or acquisition characteristics.

Building on transform-space ideas, R. Nanmaran et al. [5] investigated contrast enhancement with CLAHE followed by Discrete Cosine Transform (DCT) [6]-based fusion and then classical classifiers such as SVM and KNN for brain tumor classification. Their method uses CLAHE to expose diagnostically relevant structures, applies DCT to fuse frequency-domain coefficients from different modalities, and demonstrates that fused images can improve downstream ML classification accuracy. SVM, KNN, and decision trees were tested with fused images SVM obtained maximum accuracy of 96.8%. The work addressed the practical question of whether fusion actually improves classifier performance and provided empirical evidence of gains, but it also highlighted a limitation of such pipelines which is heavy dependence on preprocessing and feature engineering, and the limited representational power of classical classifiers compared to hierarchical feature learners.

P.K. Ambily et al. [7] applied DWT [8] to decompose images into multi-scale coefficients, combined coefficients to form a fused image, and then used a feed-forward ANN for tumor segmentation and classification. The wavelet decomposition was intended to preserve both low-frequency (structural) and high-frequency (edge/texture) details across modalities, and the ANN leveraged those fused cues for detection. This approach improved over single-modality processing by explicitly preserving multi-scale features, yet it exposed the limitation of early shallow neural approaches which is they relied on hand-tuned decomposition and shallow representations, lacking the deep hierarchical features that modern CNNs can extract, and were sensitive to segmentation heuristics and preprocessing choices.

III. CNN Based Methods

CNNs [9] are deep learning models used to efficiently analyze structured grid-like data. They operate through convolutional layers that apply learnable filters across input images to extract features like edges and textures in early layers to anatomical patterns and semantic regions in deeper layers. CNNs typically include activation functions, pooling operations for spatial downsampling, and fully connected or classification layers to interpret extracted features. Their weight-sharing design reduces parameters. This enables efficient learning even with high-dimensional image data.

CNNs are used to detect brain tumors from single-modality MRI by automatically learning discriminative features directly from the images. The MRI scans are first preprocessed and then fed into a CNN. The convolutional layers obtain low-level features like edges and textures while deeper layers learn high-level tumor-specific patterns. The network is trained using labeled MRI data to classify images (tumor vs. non-tumor) or to segment tumor regions at the pixel level. By learning spatial and intensity variations present in MRI, CNNs eliminate the need for handcrafted features. The work proposed by Neha Sharma [10] used a CNN that classified images into tumor class and non-tumor classes. The paper proposed classification of MRI scans into two classes using VGG16 (pre-trained) and achieved training accuracy of 96.5% and testing accuracy of 90%.

CNNs have transformed medical image fusion by enabling data-driven and adaptive feature extraction from multimodal sources such as MRI, CT, and PET. Traditional fusion techniques often rely on static rules or handcrafted transforms, which can introduce spatial or spectral distortions and fail to capture complex anatomical relationships. CNN-based fusion models 1 learn to extract informative features and generate activity or weight maps that emphasize important structures from each modality based on localized context. They support multiscale fusion pipelines when integrated with Gaussian or wavelet decompositions, allowing fine-grained structural details and global intensity information to be preserved. Transfer learning using pretrained models further enhances robustness when labeled medical data are limited, while Siamese architectures effectively compare and integrate complementary features from paired inputs. This study proposed by Amini et al [11] focuses on enhancing MRI-PET fusion by utilizing deep features extracted from a pretrained VGG19 network.

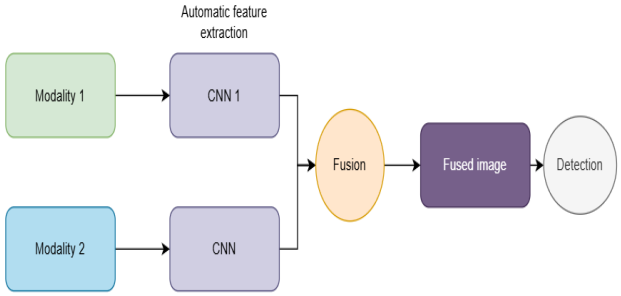

Fig. 1: Image fusion using CNNs: Features from both images are fused to obtain a combined representation.

Traditional MRI–PET fusion methods often struggle to preserve structural details of MRI and functional details of PET simultaneously. Here, the PET image is transformed to the HSI space and the intensity channel is fused with MRI after deep feature extraction. MRI and PET are processed using VGG19. The feature weights of an early layer are used to guide weighted fusion using an inner-product rule. After fusion, the image is converted to RGB, preserving structural clarity and functional contrast. The method uses a deep feature fusion strategy. The authors point out that CNN models tend to overfit and fail to capture diverse tumor characteristics, especially in small medical datasets.

Trivedi et al. [12] proposed improving fused image quality by overcoming the issues of traditional methods that rely on handcrafted rules and conventional CNN-based approaches that use fixed pooling operations. These lead to loss of important features like edges, textures, and fine structures. Since different images provide complementary structural, functional, and anatomical information, the authors proposed a fusion method using CNN with automated pooling mechanism. In this method, multiple modalities are given as input to a CNN which extracts multi-level features through convolution layers. The core innovation lies in adaptively selecting the most suitable pooling strategy based on the input features, thereby reducing spatial dimensions while preserving clinically relevant information. The automated pooling algorithm computes both max-pooled and average-pooled versions of a feature matrix and selects the final pooled output based on the difference between them. If the difference is zero, either pooling result is used; for small differences, max pooling is selected; for moderate differences, the average of max and average pooling is used; and for large differences, average pooling is chosen. The pooled features from different modalities are then fused to form a unified representation, which is reconstructed into a single fused image, resulting in improved information preservation and more reliable medical diagnosis.

Guo et al. [13] proposed a decision-level fusion network for the classification of glioma subtypes by using multiple types of MRI (T1, T1ce, T2, and FLAIR). Modality-specific DenseNet-121 models are trained on segmented tumor regions, and final predictions are obtained through weighted decision fusion. Different MRI modalities provide complementary pathological cues and that treating them independently before fusion helps preserve modality-specific strengths. The model achieves approximately 87.8% accuracy and an AUC of 0.902, outperforming both single-modality baselines and simple feature concatenation.

IV. GAN Based Methods

GANs are a class of neural networks used to generate new data that is identical to a given data. It was introduced by Ian Goodfellow [14]. GANs are used for the creation realistic images, videos, and other data formats. GANs involve two networks that fight each other, which makes the creation of highly realistic data possible. A GAN consists of two networks:

Generator: This network takes random noise as input and generates realistic data, like an image. The goal of the network is to fool discriminator into classifying the synthetic data as real. Hence it tries to minimize generator loss. During training, generator improves its parameters to reduce this loss, which improves its capability to produce realistic outputs.

Discriminator: It is a binary classifier, that tries to tell the difference between real samples of training data and synthetic samples produced by the generator. It's goal is to minimize discriminator loss. This loss reflects its ability to correctly identify real and synthetic data.

This adversarial game forces the generator to produce increasingly realistic data. The networks continuously improve at their objectives so that high quality data is produced.

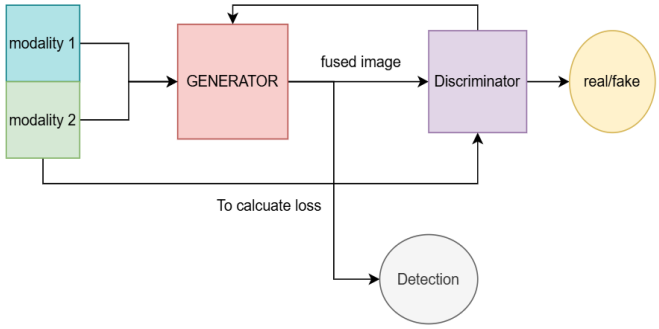

Fig. 2: Image fusion using GANs: The generator learns to create a realistic fused image that preserves information from both modalities

In studies using single modalities like MRI, GANs are used to create synthetic brain images to address the issue of limited and imbalanced medical datasets. Halima Hamid et al. [15] proposed BrainGAN, integrating GANs with CNN models to create synthetic MRI scans for improving dataset quality for classification. In this study, Vanilla GAN and DCGAN were used to produce realistic scans for data augmentation. The generated images were also used to train

CNN classifiers for brain tumor classification. The procedure involved preprocessing MRI images, training the GAN to learn the data distribution, augmenting the dataset with synthetic images, and finally applying CNNs for classification. Experimental results showed that GAN-based data augmentation significantly improved classification accuracy, robustness, and generalization performance compared to training CNNs on the original dataset alone, demonstrating the effectiveness of GANs in enhancing tumor classification. In GAN-based fusion framework 2, the generator learns to produce fused image that preserves important structural details (edges, textures) from one modality and functional or intensity information from another. The training phase uses loss functions like adversarial loss, content loss, structural similarity (SSIM), and gradient or edge preservation loss so that fused image retains relevant features from all input images.

Liu et al. [16] proposed BTMF-GAN to solve the limitations of conventional techniques. Conventional fusion techniques fail to preserve critical tumor-related features across MRI modalities. It leads to distortion and loss of relevant information. To solve this hurdle, the authors proposed a GAN-based architecture which concatenates multiple MRI modalities and feeds them to a generator to produce a fused output enriched with lesion-specific structures. The fused images are segmented and fragmented using tumor masks. Then they are evaluated by the discriminator, enabling progressive enhancement of tumor-region fidelity during training. The generator continuously improves so that it can produce tumor-preserving fused images that cannot be differentiated from input images. The method produces outputs that preserve structural consistency and detailed textures.

Safari et al. [17] proposed unsupervised GAN to fuse CT and 3D T1-Gd MRI. Traditional methods depend on rigid registration or overlays that require high computational resources and might introduce artifacts. MedFusionGAN solved these issues using an end-to-end architecture which uses multiple complementary loss functions that help to maintain anatomical correctness and clarity in texture resulting in visually realistic outputs. Preprocessing involve registration, masking, normalization, and augmentation across the GLIS-RT dataset. In evaluation, MedFusionGAN achieved a Dice score of 0.96 and fusion time of 1.9 seconds per scan.

V. Transformers

Transformers are powerful architectures that can be used for fusing multiple modalities. These models have capability to model long-range dependencies and can produce a combined representation from different medical data. Transformers‘3 use self-attention mechanisms to automatically weigh the contributions of each input based on its contextual relevance to the target task. The model uses this to capture complementary details and reduce redundancy in information. Vision transformers (ViTs) and hybrid CNN-Transformer architectures have been in use to combine radiology scans. Similarly, multimodal transformers like ClinicalBERT and Med-ViT variants allow joint reasoning over structured and unstructured data, enabling more accurate disease prediction, prognosis estimation, and treatment recommendation. Their inherent scalability, parallelization capabilities, and ability to learn hierarchical representations make transformers particularly advantageous for medical data fusion tasks, where modalities often differ in structure, resolution, and noise characteristics. As a result, transformer-based fusion frameworks are becoming central to next-generation clinical decision-support systems, offering improved performance, reliability, and interpretability over conventional DL models.

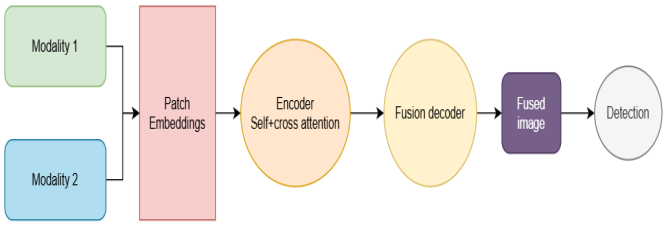

Fig. 3: Illustration of the basic image fusion process using Transformer's encoder-decoder architecture

Gómez-Guzmán et al. [18] propose an automated multi-class classification of tumors by using pre-trained CNNs and transformer architecture addressing the gap of robust, accurate, and efficient classification across four classes. in MRI, especially under class imbalance and limited data. The methodology uses the publicly available Msoud dataset (7,023 images), fine-tuning four architectures (DeiT3_base_patch16_224, Xception41, Inception_v4, Swin_Tiny_Patch4_Window7_224) with stratified 5-fold cross-validation, advanced preprocessing, data augmentation, and transfer learning. The best performing model, Swin_Tiny_Patch4_Window7_224, had 99.24% test accuracy, along with high precision, recall, F1-score, and MCC, and showed good generalization.

Dineshkumar et al. [19] proposed an Adaptive Transformer for multi-modal image fusion framework to solve the issue of loss of information, low contrast, and inefficient feature integration that commonly occur in traditional medical image-fusion methods. This approach uses adaptive attention mechanism in which attention computation is modified based on context, task, token position, or learned signals. . It dynamically weighs the importance of each modality so that the fused output has both spatial detail and broader contextual information. The method combines spatial and frequency domain features, performs preprocessing of each modality, and feeds the fused output into an object-detection module. In experiments, the system achieved strong results, achieving 98.5% sensitivity, 96.7% specificity, and F1-score of 97.2%. It demonstrated an inference speed of roughly 120 ms for each image and a model size reduction of about 35% compared to earlier fusion approaches.

Method	Publication Details	Idea	Dataset	Metrics	Limitations
IHS+PCA [3]	Changtao He et al., Procedia Engineering, 2010	Converts multimodal images to IHS space to separate intensity from functional color information. Uses histogram matching and PCA on intensity components to fuse images while preserving MRI structural detail. Recombines fused intensity with color channels, reducing color distortion but limited by hand-crafted, linear rules.	Spatial resolution of MRI and PET images were 256 x 256 and 128 x 128 pixels. All images were obtained from the Harvard medical dataset.	Used mutual information (MI) to evaluate and achieved a value of around 2.9 for each fused image which is higher compared to other methods.	Hand-crafted rules cannot learn complex nonlinear relationships between modalities and can introduce artifacts when the modalities differ greatly in noise, resolution, or acquisition characteristics.
Wavelet Transform + ANN [7]	Ambily P.K. et al., International Journal of Engineering Research and General Science, 2015.	Preprocessed images undergo multiscale decomposition using Discrete Wavelet Transform (DWT). Multiple wavelet families (Haar, Daubechies, Symlets, Coiflets, Biorthogonal, etc.) and decomposition levels are tested. An Inverse DWT is applied to the fused coefficients to reconstruct the fused image. The fused image is then passed to the segmentation stage, where thresholding based on standard deviation is used to extract abnormal (tumor) regions. For final tumor detection, a multi-layer Feed Forward Neural Network (FFNN) is employed to classify tumor and non-tumor regions.	Dataset name and size is not specified	Algorithm is evaluated on PSNR and MSE using different wavelets. Biorthagonal= 13.2296 and Symlet=13.2707 wavelet perform better on the basis of PSNR compared to other wavelet.	Heavy dependence on preprocessing and feature engineering, and the limited representational power of classical classifiers compared to hierarchical feature learners.
DCT+Machine Learning [5]	R. Nanmaran et al., Computational and Mathematical Methods in Medicine, 2022.	Discrete Cosine Transform-based fusion method was used to obtain fused images. Algorithms such as support vector machine classifier, KNN classifier, and decision tree classifiers are tested with features obtained from fused images and compared with the result obtained from individual input images.	Input images such as MRI and SPECT images were collected from Kaggle.	Performance of classifiers is measured using accuracy, precision, recall, specificity, and F1 score. SVM classifier provides an accuracy of 96.8%, the precision of 97.5%, recall of 95.12%, specificity of 97.13%, and F1 score of 96.29% for fused images.	Hand-crafted feature extraction and fusion strategies may fail to capture complex nonlinear relationships. Lack of clinical validation and interpretability.
VGG19 [11]	N Amini et al., Journal of Medical Signals & Sensors, 2022	Fused MRI and PET images with a pretrained VGG19. First, the PET image was converted from RGB space to HSI space and then features were extracted using a pretrained CNN. The weights extracted from two MRI and PET images were used to construct a fused image. Fused image was constructed with multiplied weights to images.	30 images of color PET images and 30 images of high-resolution MRI images of the brain that are registered together. All images used were obtained from the Harvard University site	Entropy, Mutual Information, Discrepancy, and Overall Performance were 3.0319, 2.3993, 3.8187, and 0.9899, respectively.	Network coefficients alone produced low-brightness fusion results. Adding two experimentally chosen constants improved quality but these constants do not generalize across different images.
MMIDFNet [13]	Shunchao Guo et al., Frontiers in Oncology,2022	MRI modalities were first segmented using a pre-trained tumor model, and tumor regions were cropped and normalized. A unified DenseNet-based network extracted features and classified images into three glioma subtypes. During inference, predictions from modality-specific models were combined using a linear weighted decision fusion strategy.	Experimental data obtained from the CPM-RadPath challenge 2020 dataset. Subtypes are Glioblastoma (“G”), Astrocytoma (“A”), and Oligodendroglioma (“O”). MRI images comprise four modalities namely T1-weighted (T1), T2-weighted (T2), post-contrast T1-weighted (T1ce), and fluid-attenuated inversion recovery (FLAIR). The cohort consisted of 221 patients collected from the original dataset, in which there were 133, 54, and 34 samples provided for subtype “G”, “A”, and “O”, respectively.	Proposed method achieved an accuracy of 0.878, an AUC of 0.902, a sensitivity of 0.772, a specificity of 0.930, a PPV of 0.862, an NPV of 0.949, and a Cohen’s Kappa of 0.773.	The proposed method has larger model size compared to many existing methods. Second, in validation, this work only considers T1, CE-T1WI, FLAIR, and T2 sequences. The performance of the algorithms is evaluated solely based on the image quality.

Method	Publication Details	Idea	Dataset	Metrics	Limitations
CNN+Automated Pooling [12]	Gargi J Trivedi et al., Indian Journal Of Science And Technology, 2022	Uses a CNN to extract multi-level features that capture both low-level details and high-level semantic information. An automated pooling layer adaptively selects the optimal pooling strategy to reduce feature dimensions while preserving important image details. The pooled features are fused and reconstructed to produce a single informative fused image.	Images obtained from Kaggle dataset and Harvard medical dataset. Sizes and number of images not mentioned in the paper.	Generates average PSNR of 36.82, average MSE of 0.53, average fusion factor of 4.62, average fusion symmetry of 0.10, average VIF of 0.91 and average processing time of 29.66seconds.	Fusion performance still depends heavily on the quality, diversity, and size of the training dataset, and increased computational complexity due to automated pooling.
BTMF-GAN [16]	Xiao Liu et al., Computers in Biology and Medicine, 2023	MRI modalities were first segmented using a pretrained tumor model, and tumor regions were cropped and normalized. A unified DenseNet-based network extracted features and classified images into three glioma subtypes. During inference, predictions from modality-specific models were combined using a linear weighted decision fusion strategy.	The BraTS 2019 dataset contained 335 cases, each with four MRI modalities (T1WI, CE-T1WI, T2WI, FLAIR) and manually annotated tumor labels (tumor core, enhanced tumor, edema, whole tumor). External clinical dataset of 10 cases with the same four modalities was used for validation, with images aligned to CE-T1WI and resampled to 240×240×16, and additional AANLIB Neoplastic Disease dataset (SPECT, GAD, T2WI) was used to test generalizability.	Outperformed existing models in metrics across both datasets, including Average Gradient, Entropy, and Mutual Information, Edge-Based Fusion Quality Index, Visual Information Fidelity, and Contrast Improvement Index. On the clinical dataset, it also leads in SSIM and PSNR.	Sample size of the dataset used was limited. Used only MRI modalities in the present study without considering other types of data. The number of subtype "G" in training set was about 60.2%, and this resulted in the class imbalance issue.
MedFusionGAN [17]	Safari et al., BMC Medical Imaging, 2023	An unsupervised generative adversarial network is used to fuse medical images by learning a direct mapping from source images to a single fused output without requiring ground-truth fusion labels. The generator is trained to preserve complementary structural and intensity information from each modality, while the discriminator enforces realism and consistency in fused image. Multiple loss components guide the training to retain anatomical structures, enhance contrast, and suppress artifacts	Publicly available multicenter medical GLIS RT dataset from the Cancer Imaging Archive con sisting 230 patients (100 males and 130 females). Modalities include 3D T1-Gd, T2-fluid-attenuated inversion recovery MRI sequences, and a CT scan The brain tumor types were glioblastoma (GBM - 198 cases), anaplastic astrocytoma (AAC - 23 cases), astrocytoma (AC - 5 cases), anaplastic oligoden droglioma (AODG - 2 cases), and oligodendroglioma (ODG - 2 case). We used 80% (11246 image slices) for training and 20% of data (2276 image slices) for testing.	Entropy = 5.2 ± 0.38, Standard Deviation = 0.44 ± 0.05, Peak Signal-to-Noise Ratio = 23.02 ± 3.5, Q (fusion quality index) = 0.64 ± 0.1, Mean Gradient = 0.20 ± 0.05, Spatial Frequency = 0.67 ± 0.14, Normalized Cross-Correlation = 0.91 ± 0.04, Mutual Information = 0.42 ± 0.29, and Structural Similarity Index Measure = 0.62 ± 0.22.	Necessitate source images to be perfectly aligned; oth erwise, fusion images may exhibit undesirable artifacts. Achieving precise alignment of medical images can be challenging and is seldom performed in diagnostic settings.
AT-MMIF [19]	Dineshkumar et al., International Journal of Computational and Experimental Science and Engineering, 2024	Multi-modal medical images are preprocessed through noise reduction and normalization. A Vision Transformer extracts global and local features, which are then fused across modalities using an adaptive attention mechanism. Finally, a Region Proposal Network detects abnormalities such as tumors or lesions in the fused image.	Dataset name and size details are not mentioned.	Framework achieved 98.1% accuracy, 98.5% sensitivity, 96.7% specificity, and 97.2% F1 score in multi-modal medical diagnosis. Inference time of 120 ms per image and 35% reduction in model size enables efficient use in resource-constrained and edge-based clinical environments.	One limitation is that the framework was tested on a predefined set of medical modalities.
FATFusion [20]	Wei Tang and Fazhi He,Information Processing & Management, 2024	FATFusion introduces a functional–anatomical Transformer-based framework for fusing PET with MRI or CT images. It employs a dual-branch Transformer with modality-specific encoders to extract complementary features. A cross-attention fusion module adaptively combines functional and anatomical information, preserving metabolic and structural details. The fused output is optimized for accurate tumor detection and medical diagnosis.	A total of 354 source image pairs were collected from the Harvard dataset. These were randomly divided into 319 training, 20 testing, and 15 validation image pairs.	mutual information (QMI = 0.8105), gradient (QG = 0.6603), phase consistency (QP = 0.7265), structural fidelity (QY = 0.9168), contrast-based quality (QCB = 0.7145), visual information fidelity (VIF = 0.3396), and MIabf = 3.5053.	FATFusion relies on paired and well-registered functional–anatomical images, which may limit applicability in real-world clinical settings. Errors in registration can negatively affect fusion quality and diagnostic reliability.

TABLE I: Comparison of methods based on ideas, datasets, and evaluation metrics

Tang and He [20] introduced FATFusion, a Functional–Anatomical Transformer–based medical image fusion framework which addresses the limitations of traditional, ConvNet-based, and existing transformer-based fusion approaches in preserving complementary information. The method tackles the issue of preserving both functional metabolic information and high-resolution anatomical details. FATFusion uses dual-branch architecture having Functional Multiscale Branch (FMB) and an Anatomical Multiscale Branch (AMB). The FMB captures metabolic and physiological patterns. The AMB focuses on preserving fine structural details and tissue boundaries. To effectively combine the complementary representations, the framework uses Functional-Guided Transformer Modules (FGTMs) and Anatomical-Guided Transformer Modules (AGTMs). These modules uses self-attention mechanism to model long-range dependencies and allow bi-directional information exchange between the two branches. Additionally, a multiscale feature extraction strategy is used along with an unsupervised loss formulation, including pixel loss and total variation loss, to improve the quality of the output. FATFusion achieved QY = 0.9002, QCB = 0.6441, VIF = 0.4185, and MIabf = 4.3946 SPECT–MRI dataset, outperforming 12 state-of-the-art methods across all seven evaluation metrics. Generalization experiments on the PET–MRI dataset demonstrated FATFusion reaching QY = 0.9168, QCB = 0.7145, VIF = 0.3396, and MIabf = 3.5053 showing strong ability to preserve functional relevance, maintain anatomical clarity, and generalize effectively.

VI. Medical Fusion Datasets

A. BraTS (Brain Tumor Segmentation Challenge Dataset)

The BraTS dataset [21] contains different types of MRI scans—T1, T1-Gd, T2, and FLAIR—from patients with glioblastoma and lower-grade glioma. It contains 2,000+ MRI volumes, with voxel-level labels for tumor sub-regions (enhancing tumor, tumor core, edema). The dataset is widely used for MRI–MRI fusion, tumor segmentation, and multimodal diagnostic modeling because of its consistent multi-sequence alignment and clinical variability.

B. TCIA

Several tumor-focused collections in The Cancer Imaging Archive (TCIA) [22] offer paired scans for oncology research, including datasets for lung cancer, head-and-neck tumors, and lymphoma. Together they include thousands of paired scans, each with tumor annotations or lesion-level metadata. These datasets are standard in fusion studies.

C. Harvard Whole Brain Atlas (Tumor Subset)

The Whole Brain Atlas [23] contains paired CT, MRI, and PET brain scans from patients with brain tumors, stroke, and neurodegenerative disorders. The tumor subset includes multiple types such as T1, T2, FLAIR, and PET sequences used to enhance visualization of brain tumors and lesion boundaries.

D. ADNI (Tumor-Relevant PET–MRI Subset)

Although primarily focused on Alzheimer's disease, ADNI [24] includes large amounts of PET–MRI data with detailed metabolic and structural imaging. Researchers frequently use PET–MRI fusion from ADNI as a benchmark for tumor-related multimodal fusion techniques due to high-quality co-registered scans, though it is not tumor-specific. Only the subsets involving structural MRI and FDG-PET are typically used for fusion models.

VII. Image Fusion Metrics

Fusion metrics are used to evaluate quality of an image produced by combining two or more source images. Their purpose is to determine how effectively the fused image preserves and enhances useful information from the originals. These metrics evaluate factors such as information content, structural clarity, edge sharpness, similarity to the source images, and the presence of noise. Fusion metrics help compare different fusion methods. It helps ensure that the fused images are reliable and informative.

A. *Entropy (ENT)*

Entropy measures the amount of detail the image contains. High value means that the fused image preserves more diagnostic information from the original modalities.

$$ENT(F) = -\sum_{i=0}^{L-1} p_i \log_2(p_i) \tag{1}$$

Where:

- p_i = probability of pixel intensity level i.
- L = Total possible intensity levels.

B. *Standard Deviation (STD)*

This metric measures the intensity variation in the fused image. Higher STD implies greater dynamic range, sharper transitions, and enhanced anatomical structure visibility.

$$STD(F) = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - \mu_F)^2} \tag{2}$$

Where:

- $F(i,j)$ = intensity value at pixel (i,j).
- μ_F = mean intensity.
- M, N = image height and image width.

C. *Peak Signal-to-Noise Ratio (PSNR)*

It calculates correctness of the final image with respect to source image. It evaluates the amount of distortion introduced by the algorithm. Higher PSNR values correspond to cleaner, less degraded fused images.

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{3}$$

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - A(i,j))^2 \tag{4}$$

Where:

- $F(i,j)$ = pixel value of fused image at (i,j).
- $A(i,j)$ = pixel value of source image at (i,j).
- MAX_I = maximum possible pixel value.
- MSE = mean squared error between F and A.

D. Mean Gradient (MG)

Mean Gradient computes average of local intensity changes. Higher MG means better preservation of structural details such as edges and ridges which are important for identifying abnormalities in medical scans.

$$MG = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{X} \tag{5}$$

$$X = \frac{(F(i+1,j) - F(i,j))^2 + (F(i,j+1) - F(i,j))^2}{2} \tag{6}$$

Where:

- $F(i,j)$ = pixel intensity at (i,j).
- $F(i+1,j)$ = pixel intensity of vertical neighbor.
- $F(i,j+1)$ = pixel intensity of horizontal neighbor.

E. Spatial Frequency (SF)

SF measures the level of detail in the final image. A higher SF means enhanced edges, improved structure visibility, and stronger contrast transitions. It is useful to assess whether fusion adds meaningful textural details.

$$SF = \sqrt{RF^2 + CF^2} \tag{7}$$

$$RF = \sqrt{\frac{1}{MN}\sum_{i,j}(F(i,j) - F(i,j-1))^2} \tag{8}$$

$$CF = \sqrt{\frac{1}{MN}\sum_{i,j}(F(i,j) - F(i-1,j))^2} \tag{9}$$

Where:

- RF = row frequency (horizontal variations).
- CF = column frequency (vertical variations).
- $F(i,j)$ = pixel intensity of fused image.

F. Normalized Cross-Correlation (NCC)

NCC calculates similarity between the final output and source images. It captures linear correlation in intensity patterns. It helps determine whether the final image retains the statistical relationships of the source modalities.

$$NCC = \frac{\sum_{i,j}(F(i,j) - \mu_F)(A(i,j) - \mu_A)}{\sqrt{\sum_{i,j}(F(i,j) - \mu_F)^2 \sum_{i,j}(A(i,j) - \mu_A)^2}} \tag{10}$$

Where:

- $F(i,j)$ = pixel value of composite image.
- $A(i,j)$ = pixel value of reference image.
- μ_F, μ_A = mean intensities of fused and reference images.

G. Mutual Information (MI)

MI quantifies the amount of content of each source image within the composite image. It evaluates the preservation of complementary anatomical and functional details.

$$MI = MI(F,A) + MI(F,B) \tag{11}$$

$$MI(F,A) = \sum_{i,j} p_{FA}(i,j)\log_2\left(\frac{p_{FA}(i,j)}{p_F(i)p_A(j)}\right) \tag{12}$$

Where:

- $p_{FA}(i,j)$ = joint probability distribution of F and source A.
- $p_F(i)$ = marginal probability distribution of the fused image.
- $p_A(j)$ = marginal distribution of source image A.

H. Structural Similarity Index (SSIM)

SSIM determines how well the structural information preserved relative to each source image.

$$SSIM(F,A) = \frac{(2\mu_F\mu_A + C_1)(2\sigma_{FA} + C_2)}{(\mu_F^2 + \mu_A^2 + C_1)(\sigma_F^2 + \sigma_A^2 + C_2)} \tag{13}$$

Where:

- μ_F, μ_A = mean intensities of fused and source images.
- σ_F^2, σ_A^2 = variances of fused and source images.
- σ_{FA} = covariance between fused image and source image.
- C_1, C_2 = small constants to stabilize the division.

VIII. Performance metrics for Classification

1. **True Positive (TP)**: A tumor image rightly classified as tumor.
2. **False Positive (FP)**: A non-tumor image wrongly classified as tumor.
3. **True Negative (TN)**: A non-tumor image rightly classified as non-tumor.
4. **False Negative (FN)**: A tumor image wrongly classified as non-tumor.

A. Precision

It tells how many of the images labeled as tumors are actually tumors, out of all images the model predicts as tumors. A high precision model minimizes false alarms.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

B. Recall

Also called True Positive Rate, it shows how many of the actual tumor images are correctly detected by the model. A high recall model is less likely to miss tumors, which is critical in medical diagnosis.

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

C. Accuracy

The metric tells how many images are correctly labelled by the model out of all the images labelled . It provides details on overall performance.

$$CorrectClassifications = TP + TN \tag{16}$$

$$TotalClassifications = TP + TN + FN + FP \tag{17}$$

$$Accuracy = \frac{CorrectClassifications}{TotalClassifications} \tag{18}$$

D. F1 Score

The F1 score is harmonic mean of precision and recall. It is used when there is class imbalance.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

E. Specificity

Also called False Positive Rate, it is the ratio of non-tumor images wrongly labelled as tumor to the total non-tumor images. It tells how often healthy cases are wrongly identified as tumors.

$$Specificity = \frac{FP}{FP + TN} \tag{20}$$

IX. Limitations and Challenges

Image fusion in medical field has many limitations because of varying characterstics of the modalities in use and the conditions under which they are acquired. Due to patient movements or distortions, there might be no consistent alignment or similar distribution between modalities. This results in unreliable outcomes. Medical images can contain noise, artifacts, and other modality-specific differences. So additional effort is necessary to capture meaningful information which makes it less practical to implement on standard clinical hardware. Scarcity of large, annotated, and multi-modal medical datasets also restricts model training, affects generalization, and leads to inaccurate outputs. Patient motion, physiological variations, and acquisition noise, introduce distortions to images that fusion models may fail to handle effectively. This causes them to saturate or generate artifacts. It requires high computational cost to train the fusion models. Color and structural correctness remain as issues because of the differences in the image content of the modalities. These differences cause inconsistencies that can challenge models trained on limited or synthetic data. Reliance on synthetic training data which does not capture real medical variability can affect the performance and limit the adoption of the model into real-world scenarios.

X. Conclusion

This paper provides a comprehensive overview of medical image fusion techniques that are designed to combine complementary information from multiple imaging modalities. It also analyzes recent deep learning-based and Transformer-based approaches and how these contribute towards producing clearer and more informative fused images. The main fusion approaches involve CNN-based models, GANs, and Transformer-architectures. These addressed problems like imbalance in contrast, noise, and other limitations associated to modalities. Medical imaging datasets and evaluation metrics were also discussed.

Although there is considerable progress in medical image fusion, some challenges still exist. Many models are trained on limited or modality-specific datasets. This limits generalization and the models may not reflect the diverse real-world clinical cases. DL–based fusion methods require huge computational resources and large datasets with precise image registration. Moreover, maintaining the structural fidelity, anatomical detail, and diagnostic relevance across different modalities continues to be an issue. For future research, it is essential to develop more robust, efficient, and generalizable fusion architectures, enhance dataset diversity, and explore real-time, hardware-friendly solutions that can better support clinical workflows and improve the practical applicability of medical image fusion techniques in real diagnostic settings.

Author contribution

Kancharagunta Kishan Babu: main idea, supervision, review of paper; Palliyana Shabarish: initial draft writing; Nakka Smitha: manuscript preparation; Preetam Pujari: formalisation and management of references and citations; Kore Abhijith: preparing comparision table; Balini Lohith Reddy: preparation of figures and formatting;

References

[1] G. Katti, S. A. Ara, and A. Shireen, "Magnetic resonance imaging (mri)–a review," *International journal of dental clinics*, vol. 3, no. 1, pp. 65–70, 2011.

[2] O. Schillaci, L. Filippi, C. Manni, and R. Santoni, "Single-photon emission computed tomography/computed tomography in brain tumors," in *Seminars in nuclear medicine*, vol. 37, pp. 34–47, Elsevier, 2007.

[3] C. He, Q. Liu, H. Li, and H. Wang, "Multimodal medical image fusion based on ihs and pca," *Procedia Engineering*, vol. 7, pp. 280–285, 2010.

[4] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[5] R. Nanmaran, S. Srimathi, G. Yamuna, S. Thanigaivel, A. Vickram, A. Priya, A. Karthick, J. Karpagam, V. Mohanavel, and M. Muhibbullah, "Investigating the role of image fusion in brain tumor classification models based on machine learning algorithm for personalized medicine," *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, p. 7137524, 2022.

[6] S. A. Khayam, "The discrete cosine transform (dct): theory and application," *Michigan State University*, vol. 114, no. 1, p. 31, 2003.

[7] P. Ambily, S. P. James, and R. R. Mohan, "Brain tumor detection using image fusion and neural network," *International Journal of Engineering Research and General Science*, vol. 3, no. 2, pp. 1383–1388, 2015.

[8] T. Edwards, "Discrete wavelet transforms: Theory and implementation," *Universidad de*, vol. 1991, no. 28-35, p. 2, 1991.

[9] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[10] A. K. Agarwal, N. Sharma, M. K. Jain, *et al.*, "Brain tumor classification using cnn," *Advances and Applications in Mathematical Sciences*, vol. 20, no. 3, pp. 397–407, 2021.
[11] N. Amini and A. Mostaar, "Deep learning approach for fusion of magnetic resonance imaging-positron emission tomography image based on extract image features using pretrained network (vgg19)," *Journal of Medical Signals & Sensors*, vol. 12, no. 1, pp. 25–31, 2022.
[12] G. J. Trivedi and R. Sanghvi, "Medical image fusion using cnn with automated pooling," *Indian Journal Of Science And Technology*, vol. 15, no. 42, pp. 2267–2274, 2022.
[13] S. Guo, L. Wang, Q. Chen, L. Wang, J. Zhang, and Y. Zhu, "Multimodal mri image decision fusion-based network for glioma classification," *Frontiers in Oncology*, vol. 12, p. 819673, 2022.
[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[15] H. H. N. Alrashedy, A. F. Almansour, D. M. Ibrahim, and M. A. A. Hammoudeh, "Braingan: brain mri image generation and classification framework using gan architectures and cnn models," *Sensors*, vol. 22, no. 11, p. 4297, 2022.
[16] X. Liu, H. Chen, C. Yao, R. Xiang, K. Zhou, P. Du, W. Liu, J. Liu, and Z. Yu, "Btmf-gan: A multi-modal mri fusion generative adversarial network for brain tumors," *Computers in Biology and Medicine*, vol. 157, p. 106769, 2023.
[17] M. Safari, A. Fatemi, and L. Archambault, "Medfusiongan: multimodal medical image fusion using an unsupervised deep generative adversarial network," *BMC Medical Imaging*, vol. 23, no. 1, p. 203, 2023.
[18] M. A. Gómez-Guzmán, L. Jiménez-Beristain, E. E. García-Guerrero, O. A. Aguirre-Castro, J. J. Esqueda-Elizondo, E. R. Ramos-Acosta, G. M. Galindo-Aldana, C. Torres-Gonzalez, and E. Inzunza-Gonzalez, "Enhanced multi-class brain tumor classification in mri using pre-trained cnns and transformer architectures," *Technologies*, vol. 13, no. 9, p. 379, 2025.
[19] R. Dineshkumar, A. Ameelia, R. Tatiraju, V. R. Kanth, and J. Nirmaladevi, "Adaptive transformer-based multi-modal image fusion for real-time medical diagnosis and object detection," *International Journal of Computational and Experimental Science and Engineering*, vol. 10, no. 4, pp. 890–897, 2024.
[20] W. Tang and F. He, "Fatfusion: A functional–anatomical transformer for medical image fusion," *Information Processing & Management*, vol. 61, no. 4, p. 103687, 2024.
[21] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.
[22] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.
[23] D. Summers, "Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 3, pp. 288–288, 2003.
[24] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.