Ingénierie des Systèmes d'Information
Vol. 31, No. 7, July, 2026, pp. 2255-2272
Journal homepage: http://iieta.org/journals/isi

Deep Learning-Based Multimodal Medical Image Fusion: A Comprehensive Review of Architectures, Fusion Strategies, Evaluation Metrics, and Future Perspectives

Noor Hasan Hassoon[1,2*], Zaihisma Binti Che Cob[3], Norziana Jamil[4], Hazim Noman Abed[5]

[1] College of Graduate Studies, Universiti Tenaga Nasional, Kajang 43000, Malaysia
[2] Department of Computer Science, College of Education for Pure Science, University of Diyala, Baqubah 32001, Iraq
[3] College of Computing and Information, Universiti Tenaga Nasional, Kajang 43000, Malaysia
[4] Department of Information Systems and Security, College of IT, UAE University, Al Ain 15551, United Arab Emirates
[5] Department of Computer Science, College of Science, University of Diyala, Baqubah 32001, Iraq

Corresponding Author Email: Noor.hason80@gmail.com

https://doi.org/10.18280/isi.310711

Received: 14 March 2026
Revised: 30 June 2026
Accepted: 15 July 2026
Available online: 31 July 2026

Keywords:
multimodal medical image fusion, deep learning, neural network architectures, fusion strategies, medical image analysis, evaluation metrics, transformer-based models

ABSTRACT

Multimodal medical image fusion integrates complementary information from different imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), to generate comprehensive representations for medical analysis. With the rapid development of deep learning (DL), numerous fusion models have been proposed; however, existing reviews often discuss architectures, fusion mechanisms, datasets, and evaluation criteria separately, making it difficult to obtain a unified understanding of current progress. This review presents a comprehensive analysis of DL-based multimodal medical image fusion methods by organizing existing approaches according to network architectures, fusion strategies, commonly used datasets, and evaluation metrics. The reviewed methods are examined from the perspectives of feature extraction capability, fusion quality, computational complexity, and clinical applicability. The analysis shows that convolutional neural networks (CNN), generative adversarial networks (GAN), transformer-based models, and emerging state-space models have become important directions for multimodal fusion. However, challenges remain regarding data availability, model generalization, computational requirements, and validation in clinical environments. This review further summarizes current limitations and discusses future research trends, including lightweight architecture, explainable artificial intelligence (XAI), hybrid learning frameworks, and clinically reliable fusion systems. The findings provide a structured reference for researchers developing efficient and robust DL-based multimodal medical image fusion methods.

1. INTRODUCTION

Medical Imaging is of vital importance for diagnosis purposes; however, information derived from a single type of medical image is not always sufficient to obtain full information about the patient's condition. This fact has prompted much scientific work on multimodal medical image fusion, which combines complementary information provided by various types of images into one informative representation. For instance, CT gives good structural information while magnetic resonance imaging (MRI) gives good contrast, and thus fusion of these images will give more valuable information. Nowadays, there has been an increasing trend towards multimodal fusion due to its ability to combine complementary information provided by various imaging techniques [1].

Individual medical images may provide incomplete or modality-specific information. An example is the difference in the focus positions, which causes objects to appear blurred on some images. Different imaging modalities capture complementary information that, when analyzed separately, may provide an incomplete representation of the underlying anatomical or functional characteristics. This challenge has drawn significant research attention [2, 3]. With the advancements in the processing of medical images, image fusion has emerged as an effective solution. Image fusion integrates complementary information from multiple source images into a single composite representation while preserving diagnostically relevant structures. Image fusion [4-6] is a specialized algorithm designed to combine multiple images into one enhanced output. It is especially in this field that multimodal medical image fusion is essential because it can be used to improve the accuracy of diagnostic and treatment planning [7].

Common medical imaging modalities include computed tomography (CT), positron emission tomography (PET), MRI, and single-photon emission computed tomography (SPECT). These modalities provide complementary anatomical, structural, functional, and metabolic information. Each imaging modality has unique features, and different sensors capture distinct imaging information of the same anatomical region. The objective of image fusion is to improve contrast,

fusion quality, and visual perception while meeting the following conditions: An effective fusion method should preserve relevant information from the source images while minimizing artificial artifacts, registration errors, and noise [8].

Multimodal image fusion assists in developing intelligent systems that can identify objects in complex conditions and make real-time decisions. For example, the fusion of CT and MRI images can support disease identification and treatment planning by integrating complementary anatomical information [9, 10]. Furthermore, multimodal imaging information enhances the decision-making of the system. Multimodal image fusion technology has wide applications in various fields, and the demand for this technology continues to increase [11]. Multimodal image fusion performs better in autonomous driving to enhance the detection of objects, decision-making, and the comprehension of scenes. Integrating data via different sensors, such as cameras and LiDAR, enables a more precise and comprehensive environmental perception [12]. This rising need for multimodal image fusion technology is currently motivating the research and development industry and academics. The increasing availability of multimodal imaging data and the demand for more accurate automated analysis have consequently intensified research on multimodal image fusion methods [13].

In recent decades, image fusion has been extensively explored [14-16] and traditional methods usually attempt to build appropriate operators to effectively extract features or components from input images. These approaches then carefully design fusion strategies to acquire coefficients that combine the extracted features from complementary modalities. Additionally, they often demonstrate poor generalization under varying conditions due to their restricted capability in constructing powerful feature representations and deriving scene-dependent fusion.

Although DL-based multimodal medical image fusion has developed rapidly, existing reviews have several important limitations. Most previous reviews provide broad overviews without establishing a unified architectural taxonomy for recently developed DL models. Moreover, the former literature tends to discuss fusion architectures and strategies, datasets, and evaluation metrics separately; thus, it is challenging to compare the approaches or comprehend their applicability in various clinical settings. These gaps indicate that more review is necessary. To overcome these limitations, recent research has shifted towards DL-based methods, which can learn the fusion process from the input data directly, leading to improved adaptability, feature representation, and robustness in multimodal image fusion. The following are the key contributions of this review:

•Offers a structured analysis of DL models used for multimodal medical image fusion, grouped by architecture type.

•Provides a general framework of classification for image fusion strategies based on DL.

•Analysis the publicly available datasets commonly used in multimodal medical image fusion, highlighting modality combinations.

•Presents and classifies the evaluation metrics used to assess fusion quality based on what aspect of image quality they assess.

•Provides a structured, critical discussion that categorizes fusion architecture, challenges, and future directions and offers clarity on strengths and limitations of models.

This study was based on a literature review of works concerning DL-based multimodal medical image fusion. Relevant publications were identified from peer-reviewed journal articles, conference proceedings, and other academic sources. References were selected based on their conceptual relevance and contributions to fusion architectures, DL models, fusion strategies, datasets, and evaluation metrics. Through this review approach, a broad range of studies was incorporated to provide a comprehensive perspective on recent advancements and open challenges in multimodal medical image fusion.

While many review studies have explored image fusion in various domains, including remote sensing, computer vision, and general biomedical imaging, this study distinctly contributes to the specific field of DL-based multimodal medical image fusion through several key differentiators. Table 1 compares the present review with existing review studies.

Table 1. Comparison with existing review papers on image fusion

Ref.	Main Objective
[5]	It summarizes recent developments of deep learning (DL)-based pixel-level image fusion methods, classifies existing methods into different strategies of fusion, propose future directions in the research of pixel-level image fusion.
[9]	Present a review on multimodal fusion methods for medical images, focusing on various common traditional and DL models, their application, evaluation metrics, and challenges to inform future work.
[17]	Compare and classify several fusion techniques based on DL through different sensor domains, identifying fusion levels, architecture, and applications outside of medical images.
[18]	It offers a comprehensive analysis of medical image modalities, fusion techniques, available multimodal image fusion datasets, and evaluation metrics. It provides a foundational guide for researchers in the fusion domain.
[19]	It summarizes medical image fusion techniques based on DL, categorizes them according to the level of image fusion and the architecture of the DL network, and highlights problems and future research directions.
[20]	It reviews multimodal medical image fusion methods, characterizes them according to image decomposition and reconstruction, fusion rules, image quality assessment, and benchmark dataset experiments, and highlights the scientific challenges in multimodal medical image fusion.
[21]	It summarizes the current algorithms of medical image fusion, classifies them into categories namely morphological, human visual system operator-based, sub-band decomposition, neural network-based, fuzzy logic-based, and discusses them in comparative manner.
[22]	It offers a comparative review of the multimodal biomedical data fusion methods based on DL, featuring architectures, applications, challenges, and future directions.
[23]	The paper categorizes the multi-focus image fusion (MFIF) algorithms based on DL and highlights the available datasets, performance metrics, existing challenges, and future research directions.
Our study	This review provides a unified analysis of DL-based multimodal medical image fusion by jointly examining network architectures, fusion strategies, datasets, and evaluation metrics. It further presents a critical comparison of existing approaches and highlights emerging research trends, open challenges, and future directions for clinically applicable fusion systems.

This review, unlike the previous ones, not only provides a comprehensive analysis of DL-based multimodal medical image fusion in terms of network architectures, but also introduces the multimodal fusion strategies, medical image databases, and evaluation metrics into one framework for a unified analysis. It also provides a structured overview of the most recent DL models, a critical comparison of their advantages and disadvantages, and an overview of the new research directions, including lightweight models, state-space models, and explainable AI. This integrated view allows a more complete picture of what is happening and what work needs to be done than previous surveys.

2. MULTIMODAL IMAGE FUSION BASED ON TRADITIONAL APPROACHES

Traditional multimodal image fusion methods comprise established mathematical, statistical, and transform-based techniques for integrating information from multiple source images. These methods normally use statistical or mathematical models and involve the extraction of features from the input datasets and integrating them to produce one output. Transform-based, dictionary-based, and statistical methods are among the most widely reported traditional fusion approaches in the literature. Various applications of medical imaging, surveillance, industrial image processing, and remote sensing have been carried out using these traditional strategies. Numerous studies have investigated traditional image fusion techniques across different application domains [24-27]. Although they have demonstrated good results, they are also limited by their high computational complexity and the lack of robustness to noise and artifacts. Consequently, recent studies have aimed to develop DL-based solutions capable of learning the fusion method directly on input data and circumventing some of the drawbacks of the traditional ones. Figure 1 illustrates the traditional approaches for multimodal image fusion [28].

Tang et al. [17] compared multimodal sensor fusion using DL approaches with traditional approaches and highlighted several advantages of deep multimodal learning for multimodal data-processing tasks. The classic early fusion methods have traditionally been associated with large data preprocessing, in which features are to be manually modeled following prior domain understanding and data specifications. These approaches typically involve explicit feature selection and dimensionality reduction to optimize data representation. Nonetheless, data-level fusion may have scalability issues when subsequent fusion rules might have to be redefined to fit changed circumstances. Nevertheless, despite such limitations, early fusion methods often need less training data and fewer hyperparameters, which is why they are computationally efficient and comparatively simpler to apply in specific applications. As a result, although conventional early fusion methods have their simplicity and efficiency, DL-based multimodal fusion generally provides greater adaptability and representational capacity, particularly for complex and large-scale data fusion tasks.

Traditional image fusion techniques are still computationally efficient and easy to interpret, but heavily dependent on manually designed feature extraction and fusion rules. These restrictions have led to the development of approaches based on DL that can learn strong representations of multimodal information directly from data.

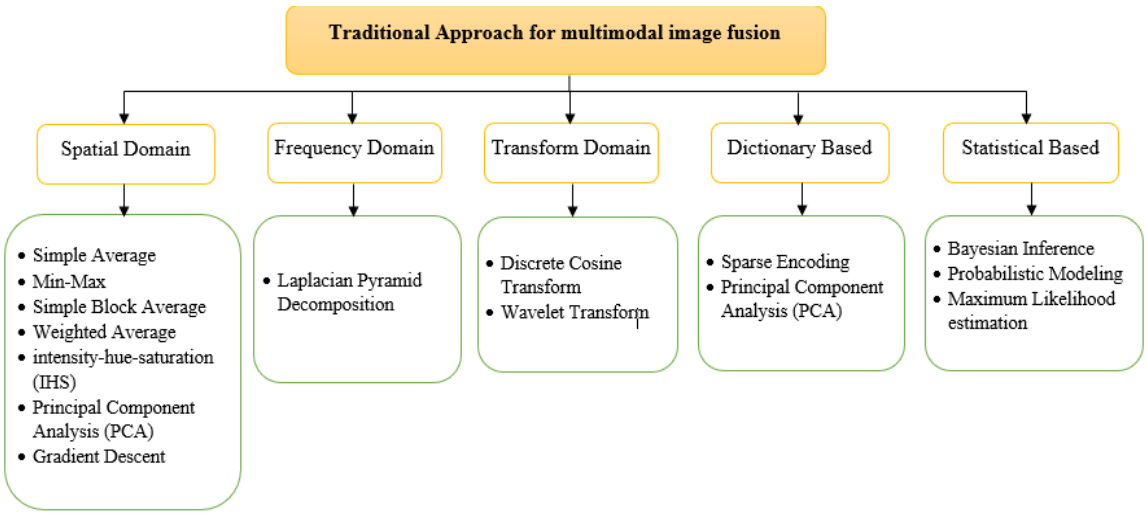

Figure 1. Traditional approaches for multimodal image fusion

3. MULTIMODAL IMAGE FUSION BASED ON DEEP LEARNING APPROACHES

Through DL, the learning of multilevel structure representations of data can be performed through computational models incorporating several layers of processing [29]. The adoption of DL for image fusion has been primarily motivated by the limitations of traditional approaches. The popularity of DL as a tool in image fusion applications is related to its capability of acquiring complex intermodal mapping automatically and its ability to process large-scale datasets efficiently [7]. The classification presented in this review is based primarily on the dominant DL architecture, as it reflects the fundamental design principles and learning mechanisms of each model. Some proposals include more than one architecture (e.g., CNN–Transformer or CNN–GAN models), but they are classified by their main architecture. In Section 4, fusion strategies are considered separately to give a complementary view on how multimodal information is integrated. Table 2 summarizes the classification of the reviewed DL models.

Table 2. Classification of image fusion based on deep learning (DL) models

DL Model	Key Models	Advantages	Challenges	DL Method	Ref.
Convolutional Neural Networks (CNNs)	M4FNet	Strong feature extraction	Computational complexity, high model training time	Unsupervised	[30]
	Image fusion Convolutional Neural Network (IFCNN)	Efficient for high-resolution images	Computational complexity, lack of real-time performance	Supervised	[31]
	CNN-Based Multi-Scale Transformation	Pre-trained architectures available	Computational complexity	Supervised	[32]
	CNN-VGG19	Can handle multi-scale features	Computational complexity	Supervised	[33]
	Siamese CNN	Works well with feature-level fusion	Computational complexity, long training time	Supervised	[34]
	Multi-Objective Differential Evolution CNN	Effective feature extraction and fusion using DBN with fuzzy logic	High computational cost, long training time, limited benchmark datasets for validation	Unsupervised	[35]
	CNN-Based Multi-Scale Feature Fusion	Accurate feature-level fusion	High computational cost, limited generalization across modalities	Supervised	[36]
	CNN-Based Weighted Feature Fusion	Effectively preserves structural details and image quality	High computational requires more training data	Unsupervised	[37]
	Deep CNN with Contrast Enhancement	Effective feature-level fusion and contrast enhancement.	High computational, risk of overfitting	Supervised for Feature Extraction, Unsupervised for Contrast Enhancement	[38]
	Multi-Scale Residual Attention CNN	Effectively enhances fusion quality	High computational cost, sensitivity to noise, and risk of overfitting	Supervised	[39]
Generative Adversarial Networks (GANs)	EMFusion	Enhance structural fidelity without requiring paired data.	Computational complexity, training instability	Unsupervised	[40]
	DCNN	Enhance structural detail and multimodal information quality	High computational cost, long training time	Supervised	[41]
	MedFusionGAN	High-quality fused images	Training instability, dataset limitation	Unsupervised	[42]
	Dense Block-Based GANs	Minimal artifacts	High computational cost, training instability, overfitting in small datasets	Unsupervised	[43]
Hybrid Architectures	CNN-Transformer Hybrid	Superior fusion by combining multiple models	Increased model complexity	Supervised	[44]
	Laplacian Pyramid CNN-Transformer	Preserves fine details	Computational complexity, training stability, parameter tuning for various modalities	Supervised	[45]
	CNN-GAN Hybrid for Image Fusion	Can be optimized for real-time applications	Computational overhead, limited dataset availability	Supervised for Feature Extraction and Unsupervised for Optimization	[46]
	Transfer Learning with VGG19 and ResNet-101	Leverages pre-trained models for improved performance	High computational cost, training instability	Supervised	[47]
	Multi-Layer Feature Concatenation CNN	Enhance brightness, contrast, and feature retention in fused images.	High computational cost, long training time, and generalization challenges to different modalities	Supervised	[48]
	CNN-Pyramid-Based Fusion	Allowing accurate reconstruction of structural and intensity details.	Computational complexity, extended training time	Unsupervised	[49]
	Unified CNN-Based Image Fusion	Achieves high-quality in segmentation and diagnosis	High computational, dataset limitations	Unsupervised	[50]
	VGG19 and CNN	Enhances fusion clarity	High computational, dependency on high-quality datasets	Supervised	[51]
	Multibranch CNN with Semantic Perception & Attention Mechanisms	Enhances semantic understanding and detail retention	High computational, unstable training	Supervised	[52]
Deep Belief Networks (DBNs)	DBN-Based Fusion, DBN with Fuzzy-Based Fusion	Effective unsupervised Feature Extraction; can learn representations without labeled data	Computationally expensive, difficult to train, less commonly used today	Supervised	[53]
State Space Models (Mamba)	FusionMamba	Adaptive feature fusion, Efficient for real-time medical applications, works well for multi-modal fusion tasks	Novel approach, Limited benchmarking, requires extensive validation	Unsupervised	[54]

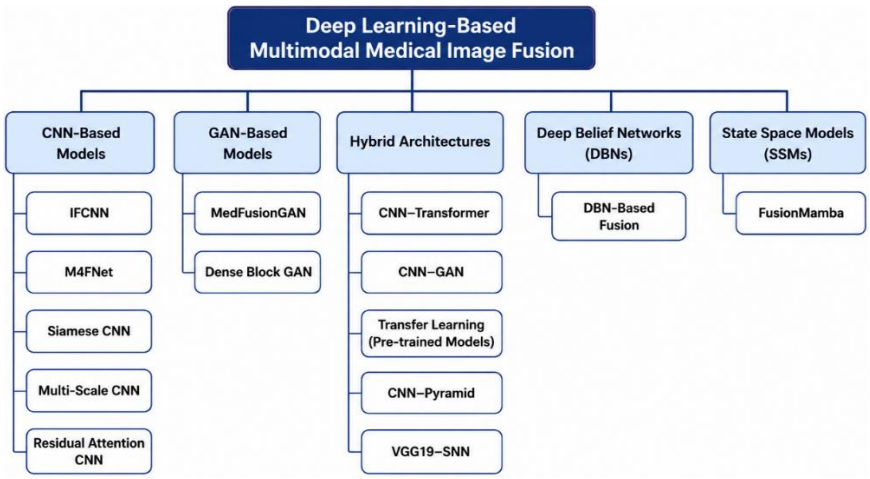

Figure 2. Taxonomy of deep learning (DL) architectures for multimodal medical image fusion

Table 2 summarizes the key characteristics, advantages, limitations, and learning paradigms of the reviewed DL models, while Figure 2 presents their architectural taxonomy.

3.1 Convolutional Neural Networks

A CNN is a class of deep neural network that employs convolution operations to process spatially structured input data. The CNN is one of the most commonly used DL architectures used to analyze spatial patterns [55]. CNNs learn spatial characteristics such as edges, corners, textures, and more abstract features that characterize the target object. The learning process consists of successive convolution operations that progressively extract hierarchical feature representations. Pooling operations reduce the spatial dimensions of the feature maps while retaining relevant information, enabling the aggregation of low-level features into higher-level representations. Similar to other neural network models, a CNN consists of neurons organized into successive layers [56].

More recently, DL approaches, particularly CNNs, have substantially advanced image fusion by enabling automated hierarchical feature extraction. Ding et al. [30] offered a framework for multimodal fusion that preserves multi-scale and multi-receptive field features, incorporating hybrid dilated convolution and an attention-based fusion strategy to retain structural details and improve diagnostic imaging. Zhang et al. [31] suggested IFCNN which is a fully convolutional DL framework designed for fusing various types of images, achieving high generalization ability without requiring fine-tuning. Xia et al. [32] presented a DL-based image fusion framework that integrates CNN and multi-scale transformation, adaptively extracting high- and low-frequency components for fusion. Elzeki et al. [33] applied a two-layer fusion approach combining Nonsubsampled Contourlet Transform (NSCT) and CNN-VGG19 to enhance CXR COVID-19 images for improved feature extraction and analysis. Liu et al. [34] proposed a novel CNN-based fusion technique that generates a weight map from two input images, decomposes them using Laplacian pyramids, and adaptively fuses features to enhance multimodal medical imaging. Kaur and Singh [35] used a DL-based image fusion framework combining NSCT decomposition, feature extraction using Xception CNN, and feature selection via multi-objective differential evolution. The proposed framework achieved high-contrast and high-quality fused images with optimal feature retention. Li et al. [36] proposed a CNN-based supervised image fusion framework that integrated deep feature learning with multimodal medical imaging. The model used training datasets to learn optimal fusion weights. Kahol and Bhatnagar [37] proposed an unsupervised DL architecture using Siamese CNNs for feature extraction and fusion, aiming at improving medical image quality and preserving structural details. Bhutto et al. [38] proposed a deep CNN-based fusion approach integrating anisotropic diffusion, PCA, and novel contrast enhancement strategies. Li et al. [39] proposed a double-branch CNN-based fusion approach integrating residual attention blocks and multiscale convolutional mechanisms. Xu and Ma [40] proposed an unsupervised medical image fusion model that combined surface- and deep-level constraints to improve structural detail and visual fidelity. The model achieved competitive performance across multiple modalities (CT, MRI, PET, and SPECT) without requiring paired training data. Liang [41] suggested a novel DL-based fusion framework using DCNNs to extract features from decomposed principal components, followed by a weighted fusion approach and sum rule for detail preservation.

In summary, CNN-based methods are the most widely used architectures for multimodal medical image fusion because of their powerful feature extraction ability and stable structure preservation. Supervised models are typically more accurate in fusion tasks, while unsupervised models are more flexible in situations where labeled data are scarce. However, their high computational cost and restricted ability to acquire long-range dependencies continue to motivate the development of more sophisticated architectures.

3.2 Generative Adversarial Networks

GANs enable the learning of deep representations without requiring extensive annotated training data. This learning process is achieved through adversarial optimization between a generator and a discriminator. GANs can learn representations that are applicable in a number of areas such as image synthesis, style transfer, image super-resolution, semantic image editing, and classification [57]. Safari et al. [42] introduced MedFusionGAN, an unsupervised GAN framework for fusing MRI and CT images. The framework enhanced tumor delineation and reduced radiotherapy planning time. Zhao et al. [43] proposed a novel fusion method integrating dense blocks for feature extraction and deep convolutional GANs for adversarial image synthesis.

In comparison to CNN-based techniques, GANs generate more realistic fused images and preserve anatomical details via adversarial learning. However, their practical applicability is constrained by training instability, high computational cost, and sensitivity to hyperparameter selection.

3.3 Hybrid architectures

Hybrid architectures in DL combine different models or techniques to leverage their strengths and overcome individual limitations. These architectures are employed to enhance efficiency, generalization, and performance in tasks including image fusion, classification, segmentation, and generation. CNNs, Transformers, GANs, low-rank models, or pyramids are commonly combined in hybrid models to extract, transform, and combine information in an effective way. Hybrid architectures have become increasingly relevant in DL-based image fusion because they integrate complementary modeling capabilities to improve feature representation, generalization, and fusion performance. Different hybrid configurations address specific methodological requirements: CNN–Transformer architectures integrate local and global feature modeling, CNN–GAN architectures combine feature extraction with adversarial image generation, and low-rank hybrid approaches aim to improve computational efficiency.

Luo et al. [44] suggested LapH, a hybrid CNN-transformer architecture that combined local feature extraction using CNNs with global feature correlation using Transformers. The architecture applied Laplacian Pyramid decomposition to process images at multiple resolutions. To enhance accuracy in detecting Alzheimer's, Beatrice et al. [46] suggested a multimodal DL system combining a CNN to extract features and a GAN to sharpen fused images, as a way of enhancing the system's accuracy. Wang et al. [45] proposed a CNN-based method utilizing Siamese networks for weight map generation and contrast pyramid decomposition for multi-scale fusion. Do et al. [47] introduced a transfer learning-based feature extraction model (TL_VGG19) with an adaptive fusion strategy based on the Equilibrium Optimization Algorithm (EOA). Dinh [48] combined Bilateral Texture Filtering (BTF) and transfer learning with ResNet-101 for medical image fusion. The method used COA to optimize texture fusion and prevent the loss of brightness and contrast, thereby enhancing feature retention in the fused images. Liang et al. [49] presented a CNN-based end-to-end fusion network that integrates feature extraction, feature fusion, and image reconstruction without complex manual rules. Balasubramaniam et al. [50] proposed a unified CNN-based fusion model integrating IFCNN and U2Fusion for medical image fusion, focusing on brain tumor segmentation and diagnosis. Allapakam and Karuna [51] proposed a hybrid DL model combining pre-trained VGG-19 and non-pre-trained SNN networks. The model used Dual-Tree Complex Wavelet Transform (DTCWT) decomposition and a stacking ensemble approach to improve image clarity. Lin et al. [52] proposed a novel CNN-based fusion method leveraging unsupervised image segmentation for improved semantic understanding and image detail retention.

Hybrid architectures have the advantage of fusing features from multiple models of DL and, as a result, improve feature representation and fusion quality across imaging modalities. Though these models tend to be both more powerful and more complex than single-model approaches, they are also typically more compute-intensive.

3.4 Deep Belief Networks

DBNs are a type of DL architecture composed of several layers of probabilistic models, that is, Restricted Boltzmann Machines (RBMs) or Autoencoders stacked on each another. DBNs refer to unsupervised types of neural networks utilized in feature learning, dimensionality reduction, and classification processes. Kaur and Singh [53] proposed a novel DBN-based medical image fusion model integrating feature selection and fuzzy logic-based decision rules. The model used pre-trained RBMs for feature extraction.

DBNs are effective for unsupervised feature learning and can be applied when labeled data are limited. They are, however, less scalable and have higher training complexity compared with CNN- and Transformer-based models and have been less frequently adopted in recent medical image fusion studies.

3.5 State Space Models

State Space Models (SSMs) are a class of serial models that represent time-series or sequential data using a hidden state representation. These models describe how an internal (hidden) state evolves over time and how observations (outputs) are generated from this state. SSMs are widely used in signal processing, control systems, natural language processing (NLP), and time-series forecasting. Traditional SSMs include models such as Kalman Filters, Hidden Markov Models (HMMs), and Linear Dynamical Systems. Mamba uses a structured state-space representation to process sequential data, offering advantages in efficiency and scalability over Transformers. Xie et al. [54] integrated the Mamba state-space model with feature enhancement techniques to improve multimodal image fusion performance. The model dynamically extracted and fused texture, structure, and correlation features from different modalities.

SSMs represent an emerging direction in medical image fusion that achieves efficient modeling of long-range dependencies with low computational cost compared to transformer-based methods. However, broader validation across diverse clinical imaging datasets is required to establish their robustness and generalizability.

Overall, CNNs are widely used due to their excellent feature extraction ability, while GANs have been mainly applied to enhance the visual quality of fused images. Improved local information was incorporated with global information through complementary learning mechanisms in hybrid architectures, and emerging SSMs are presented for enhanced computational efficiency without sacrificing fusion performance. This evolution highlights a movement in research priorities from improved feature extraction and image quality to further improvements in the design of architectures that maximizes accuracy, efficiency, robustness, and clinical applicability.

4. MEDICAL IMAGE FUSION STRATEGIES IN DEEP LEARNING-BASED

Fusion strategies play a crucial role in multimodal medical image fusion as they define how information from different imaging modalities is combined to generate a comprehensive and high-quality fused image. Table 3 summarizes the major DL-based fusion strategies identified in the reviewed literature.

Table 3. Classification of image fusion strategies based on deep learning (DL)

Fusion Mechanism	Key Advantages	Identified Challenges	Ref.
Transform-Based Fusion			
A two-layer fusion approach integrating NSCT decomposition with HVS-guided perceptual fusion.	Enhances visual contrast and perceptual detail while preserving edge and texture structures.	Requires manual threshold selection; sensitive to NSCT parameter settings, affecting reproducibility and adaptability.	[33]
Uses a bilateral texture filter to split the source images into base and detail layers, then fuses with multi-resolution norms and saliency maps.	Preserves visual detail and reduces noise; avoids excessive enhancement artifacts common in intensity-based methods.	Lacks adaptive learning; rule-based decision strategy may not generalize well across diverse modality combinations.	[48]
Feature-Based Fusion			
Employs a fully convolutional encoder–fusion–decoder framework with dense feature extraction from multimodal inputs, followed by an average fusion rule.	Demonstrates general applicability across modalities with efficient training and implementation.	Lacks inter-modal relationship modeling; employs a manually defined fusion rule, which limits adaptability.	[31]
Utilizes a supervised CNN to extract spatial and spectral features from registered image blocks and perform learned fusion.	Yields strong visual quality and metric performance across CT, MRI, and SPECT image sets.	Rely on careful preprocessing and slightly underperforms in standard deviation and relative entropy metrics.	[36]
Optimization-Based Fusion			
Applies multi-objective optimization with convolutional sparse representation and a CNN-based framework for fusion rule selection.	Enables noise suppression and optimized detail retention using objective-driven selection.	Involves parameter tuning and increased complexity for multi-objective balancing.	[47]
Combines NSCT-based decomposition and Xception features, with optimal feature selection through multi-objective differential progress.	Integrates deep feature learning with evolutionary selection to improve robustness and fidelity.	Requires extensive computational resources and careful parameter tuning for alignment.	[53]
Generative Model-Based Fusion			
Uses DBN in combination with fuzzy logic to classify and fuse informative image regions.	Achieves high entropy and mutual information with strong edge preservation and reduced visual artifacts.	Demands multi-stage training and tuning of fuzzy rules; limited validation across diverse modalities.	[35]
Employs an unsupervised GAN with a PatchGAN discriminator, combining content, SSIM, and perceptual losses for image-level fusion.	Preserves both bone and soft tissue contrasts with high visual fidelity and minimal distortion.	Requires aligned input images; training instability and mode collapse risks in GAN framework.	[42]
Combines DenseNet-based encoder-decoder architecture with deep convolutional GAN and Lmax-norm fusion rule.	Effectively integrates anatomical and functional information while reducing manual tuning.	May overemphasize intensity in CT, leading to reduced soft tissue visibility and slower inference speed.	[43]
Utilizes a DBM for supervised block-wise fusion of aligned multimodal medical images.	Automates the fusion pipeline and supports batch processing with strong structural retention.	Performance depends on registration accuracy and large labeled datasets.	[7]
Hybrid Fusion Strategy			
Combines pyramid decomposition with CNN-based activity level measurement and local similarity-based fusion rule.	Improves perceptual quality and structural clarity by jointly optimizing decomposition and fusion weights.	Requires extensive training data and is sensitive to similarity thresholds; relatively high computational cost.	[34]
Incorporates a multi-branch, multi-scale CNN with semantic segmentation-guided loss and attention mechanisms.	Captures rich semantic features and preserves both global and local structure in fused outputs.	Complex training pipeline with high computational requirements; performance depends on segmentation reliability.	[52]
Utilizes stacked convolutional layers with multi-scale Gaussian kernel initialization and autoencoder-style architecture.	Performs efficient decomposition and preserves fine image details without predefined filters.	Manual fusion rule design and training difficulty due to model depth.	[32]
Combines NSCT and FFT decompositions with PCNN-based feature selection, fuzzy logic fusion, and FGPCNN classification.	Achieves high accuracy for Alzheimer detection; robust to modality variations.	Complex pipeline with reliance on domain-specific knowledge and FFT limitations in spatial representation.	[46]
Employs an ensemble of Siamese Neural Networks and pretrained VGG-19, fused via DTCWT.	Leverages both local and global information with strong quantitative and visual performance.	High computational complexity; limited to pairwise modality fusion.	[51]
Unsupervised Siamese CNN trained with MS-SSIM-based loss and multi-scale skip connections.	Minimizes training needs and achieves effective texture and contrast fusion.	Slight drop in mutual information performance; minor visual artifacts under certain conditions.	[37]
Integrates surface-level saliency constraints and deep-level feature uniqueness using unsupervised encoder-decoder CNN blocks.	Reduces mosaics, retains chromatic and spatial fidelity across modalities.	Multi-phase training requires parameter tuning and fusion strategy balancing.	[40]

The nonlinear anisotropic diffusion method is used to divide images into main parts and detailed parts.	Fast convergence with high visual clarity, particularly effective for CT and MRI combinations.	Limited validation across other modality pairs and clinical scenarios.	[38]
Employs a Mamba-based state-space module with DVSS blocks and a UNet backbone for dynamic feature enhancement.	Achieves state-of-the-art performance with efficient global-local representation and low latency.	Model complexity and coordination among modules demand precise tuning and system integration.	[54]
Integrate multi-receptive-field CNN modules and multi-scale wavelet-based feature extractors with attention-aware fusion.	Preserves semantic structure, enhances detail and segmentation robustness.	Fixed input size limits adaptability to diverse resolution inputs.	[30]
Applies dual-branch CNN encoding followed by multi-layer concatenation and up-convolutional decoding.	Maintains both low- and high-frequency features; structurally stable and computationally efficient.	Not easily extendable to more than two modalities; structurally fixed architecture.	[49]
Unifies IFCNN-based feature fusion, U²Fusion's adaptive weighting, and RP-Net segmentation for joint fusion.	Ensures high semantic fidelity and segmentation-aware fusion with general robustness.	Evaluated only on limited modality types; broader generalization needs further validation.	[50]
Fuses LatLRR-decomposed principal components using VGG-16 and salient regions via summation.	Combines global structure with fine texture preservation across modalities.	Depends on aligned inputs and static feature extraction from pretrained VGG-16.	[41]
Combines CNN-based high-frequency branch with transformer-based low-frequency fusion through Laplacian pyramid.	Balances texture sharpness and contextual completeness; supports variable input sizes.	May result in dim outputs under low-light input; complex model design.	[44]
Employs a Siamese CNN for weight map generation, followed by contrast pyramid and region-based adaptive fusion.	Preserves contrast and structure; effective across CT, MRI, PET, and SPECT.	Requires accurate registration; multi-stage processing adds computational overhead.	[45]

Note: NSCT = Nonsubsampled Contourlet Transform, DBN = Deep Belief Networks, DBM = Deep Boltzmann Machine, GAN = Generative Adversarial Network, CNN = Convolutional Neural Network, CT = computed tomography, MRI = magnetic resonance imaging, PET = positron emission tomography, SPECT = Single Photon Emission Computed Tomography, DTCWT = Dual-Tree Complex Wavelet Transform.

4.1 Transform-based fusion strategies

Transform-based fusion methods such as NSCT and bilateral filtering focus on enhancing edge and frequency-domain detail. Elzeki et al. [33] offered a two-layer NSCT-based fusion integrated with a perceptual model inspired by the Human Visual System (HVS), demonstrating improved contrast and saliency retention. In contrast, Dinh [48] utilized a bilateral texture filter along with multi-resolution norms, which preserved fine visual details and demonstrated robustness to noise but lacked adaptive learning capabilities. Both methods are inherently rule-based and sensitive to manual parameter tuning, which limits scalability to diverse imaging conditions.

4.2 Feature-based fusion strategies

Zhang et al. [31] proposed an encoder-fusion-decoder framework with an IFCNN, which learned efficient dense-feature representations from multimodal inputs but employed a fixed averaging fusion rule, limiting its flexibility in modeling complex relationships between modalities. Li et al. [36] continued by using this method to train a supervised CNN with registration of blocks of images, which demonstrated greater adaptability for the fusion of CT, MRI, and SPECT images compared with unsupervised CNN training, but did not perform as well in some statistical measures including the relative entropy and standard deviation.

4.3 Optimization-based fusion strategies

Do et al. [47] proposed a convolutional sparse representation-based fusion model incorporating multi-objective optimization for feature selection under multiple constraints. Kaur and Singh [53] integrated NSCT-based decomposition and Xception-derived features with multi-objective differential evolution for optimal feature selection. Both approaches demonstrated effective detail preservation; however, they required substantial computational resources and extensive parameter tuning to maintain robustness across imaging modalities.

4.4 Generative model-based fusion strategies

Generative models have gained significant popularity due to their ability to learn data distributions. Kaur and Singh [35] proposed a DBN-based model integrated with fuzzy logic for block-level fusion, achieving high entropy and reduced visual artifacts but requiring complex multi-stage training. Safari et al. [42] introduced MedFusionGAN, an unsupervised generative adversarial model that achieved excellent visual quality and preserved both soft and hard tissue information, albeit with challenges in training stability and input alignment.

Zhao et al. [43] combined Dense Block CNNs with a deep convolutional GAN architecture using an Lmax-norm fusion rule, effectively balancing anatomical and structural detail, though some CT regions were overemphasized. Li et al. [7] developed a Deep Boltzmann Machine (DBM)-based model, which enabled batch fusion of CT, MRI, and SPECT with structural consistency, but it showed sensitivity to registration and a dependence on large-scale labeled data.

4.5 Hybrid fusion strategies

Recent advances are dominated by hybrid approaches that incorporate several techniques. Liu et al. [34] applied pyramid decomposition and CNN-based similarity-based fusion, which improves perceptual fine-detail preservation at the cost of increased computational complexity. Lin et al. [52] improved this by combining multi-branch CNNs with semantic segmentation and attention mechanisms, achieving greater structural and semantic consistency while requiring intensive

training. Xia et al. [32] demonstrated strong structural fidelity and efficient inference, whereas the framework proposed by Beatrice et al. [46] achieved high classification performance but relied more extensively on domain-specific processing components.

Allapakam and Karuna [51] used a combination of Siamese Neural Networks and VGG-19 features, where DTCWT fusion was used and only pairwise fusion was possible. Kahol and Bhatnagar [37] introduced the idea of an unsupervised Siamese CNN that applies the loss based on the MS-SSIM, which reduces training complexity and preserves texture, although lower mutual-information performance was reported under some conditions. Xu and Ma [40] proposed EMFusion, which combined surface-level saliency and deep uniqueness constraints using unsupervised CNN blocks and demonstrated strong chromatic and spatial retention with fewer mosaic artifacts. Bhutto et al. [38] employed deep CNN-Based feature extraction, anisotropic diffusion for image decomposition, and principal component analysis for base fusion.

More recent hybrid and advanced architectures further extend these fusion mechanisms. Xie et al. [54] introduced FusionMamba, a state-space model-based system, using DVSS blocks and a U-Net backbone, which provided improved dynamic feature representation and low-latency functionality. FusionMamba demonstrated broader generalization, whereas M4FNet, proposed by Ding et al. [30], combined multi-receptive-field CNN features, wavelet-based fusion, and attention mechanisms. Liang et al. [49] proposed MCFNet, which involves the use of two-branch CNN encoders and multi-layered concatenation with up-convolutional decoding. Balasubramaniam et al. [50] further expanded fusion capabilities by unifying IFCNN and U²Fusion with RP-Net for segmentation-guided adaptive fusion, achieving higher semantic fidelity but only validated on limited modalities.

Liang [41] fused LatLRR-decomposed components using a pretrained VGG-16 network and summation rules, preserving both global and local features, though constrained by static feature extractors. Luo et al. [44] presented LapH, a Laplacian pyramid hybrid model combining CNN and transformer branches, allowing flexible input sizes with enhanced texture and context preservation, but introducing illumination bias. Wang et al. [45] utilized a contrast pyramid with a Siamese CNN for region-based adaptive fusion, effective across CT, MRI, PET, and SPECT but sensitive to registration accuracy and complex processing stages.

Overall, no fusion strategy is universally optimal across all multimodal medical imaging scenarios. The selection of a suitable strategy depends on imaging modalities, medical objectives, computational conditions, and availability of training data, highlighting the need for application-specific fusion frameworks.

5. DATASET

To support the systematic evaluation of medical image fusion methods, Table 4 summarizes the publicly available datasets reported in the reviewed studies. The datasets are medical imaging datasets, but some computer vision datasets (e.g., ImageNet and KAIST) are included because they are used for pre-training, transfer learning, or testing DL models for multimodal medical image fusion.

The reviewed datasets vary substantially in imaging modality, anatomical region, and clinical application. For a reliable evaluation and meaningful comparison of DL-based fusion methods, it is crucial to choose datasets for the intended fusion task that are as similar as possible.

Table 4. Medical image fusion datasets

Dataset	Year	Modality	Body Organ	Target Disease/Application	Address
Harvard Medical School Whole-Brain Atlas	1999	MRI, CT, PET, SPECT	Brain	Normal brain anatomy, aging, stroke, brain tumors, Alzheimer's disease, multiple sclerosis, and other neurological disorders	https://www.med.harvard.edu/aanlib/
GLIS-RT Cancer Imaging Archive	2021	CT, REG, RTSTRUCT, MRI	Brain	Glioblastoma, Astrocytoma, Low Grade Glioma	https://www.cancerimagingarchive.net/collection/glis-rt/
COVID-19 chest x-ray	2020	X-RAY, CT	Lung / Chest	COVID-19 and other pulmonary infection detection, including SARS, MERS, ARDS, and pneumonia	https://www.kaggle.com/datasets/bachrr/covid-chest-xray
Alzheimer's Disease Neuroimaging Initiative (ADNI)	2003	MRI, PET, FMRI	Brain	Alzheimer's disease, mild cognitive impairment, disease progression, and biomarker validation	https://adni.loni.usc.edu/
open access series of imaging studies (OASIS)	2010	MRI, PET	Brain	Normal aging, cognitive decline, dementia, and Alzheimer's disease	https://www.oasis-brains.org/
Medical image dataset annotation service (MIDAS)	2010	MRI, CT, PET, US, SPECT	Multiple organs	Biomedical image-analysis algorithm development, validation, segmentation, and reproducible research	https://insight-journal.org/
Image Fusion Dataset (IFD)	2017	CT, MRI	Brain	Multimodal CT–MRI image fusion benchmarking, evaluation of medical image fusion algorithms, and quantitative/visual comparison of fusion methods	http://www.imagefusion.org
BraTS2020 Brain Tumor Dataset	2020	MRI	Brain	Glioma and brain-tumor subregion segmentation	https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation/discussion
Brain MRI segmentation	2019	MRI	Brain	Lower-grade glioma detection and segmentation	https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation?rvi=1

Gastrointestinal stromal tumors (GISTs)	2023	CT, PET	Gastrointestinal tract	Gastrointestinal stromal tumor analysis and CT–PET fusion experiments	https://github.com/praneethMohan/GIST-CT-PET
Brain MRI Image	2023	MRI	Brain	Brain MRI classification, commonly including tumor-related classification	https://www.kaggle.com/datasets/ashfakyeafi/brain-mri-images/

Note: NSCT = Nonsubsampled Contourlet Transform, DBN = Deep Belief Networks, DBM = Deep Boltzmann Machine, GAN = Generative Adversarial Network, CNN = Convolutional Neural Network, CT = computed tomography, MRI = magnetic resonance imaging, PET = positron emission tomography, SPECT = Single Photon Emission Computed Tomography, DTCWT = Dual-Tree Complex Wavelet Transform.

6. EVALUATION METRICS

The evaluation of multimodal medical image fusion involves both image-quality measures and, when downstream clinical tasks are considered, conventional classification or segmentation performance metrics. In the case of medical classification tasks, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are usually employed as indicators for measuring performance of multimodal fusion methods and DL networks. One can compute a range of performance metrics, including sensitivity, specificity, accuracy, precision, F1-score, and others based on these indicators [58]. Table 5 classifies the evaluation metrics used for multimodal medical image fusion.

Table 5. Evaluation metrics of multimodal fused image

Mation-Theoretic Metrics					
Ref.	**Metrics**	**Full Name**	**Description**	**Expression**	**Desired VALUE**
[59]	MI	Mutual Information	It is applied in quantifying the similarity of the image intensity between the reference and the fused image. Good quality of the image is represented by a high value of MI.	$MI(A,B) = \sum_{a\in A}\sum_{b\in B} p(a,b) log\,\langle \frac{p(a,b)}{p(a)p(b)} \rangle$	Higher value
[60]	Q_{MI}	Normalized Mutual Information	That is the one which calculates the mutual information between a fused image and the one that is provided as the source to be fused. Those are the coefficients of the correlation between fused images and source images.	$Q_{MI} = 2\left[\frac{MI(A,F)}{H(A)+H(F)} + \frac{MI(B,F)}{H(B)+H(F)}\right]$	Higher value
[61]	FMI	Fusion Mutual Information	It is applied to calculate the dependence level between input images and fusion images. An FMI value higher denotes a better fused image quality.	$FMI = MI_{IpIf} + MI_{ImIf}$	Higher Value
[37]	QFMI	Quality Feature Mutual Information	It is employed to compute the similarity of the reference image and the fused image. It is applied to measure the quality of the fused image, and it is also used to ensure that essential features of the reference images are not lost in the fused image.	A weighted or quality-focused variant of FMI—specific formula depends on implementation.	Higher value
[62]	Q_{PC}	Phase congruency	Is an edge and feature-saving measure applied to image fusion, comparing the preservation of important structures in the fused image to the source images.	$PC(x)\frac{\sum_n W_n(x)A_n \cos(\emptyset_n(x) - \emptyset(x)) - T}{\sum_n A_n(x) + \epsilon}$	Higher value
[9]	$Q^{AB/F}$	Quality Assessment Based on Gradient	Computes the value of edge information of the source image to the fused image in the form of Sobel edge detector operators. The higher the $Q^{AB/F}$ ratio is, the more information is transferred out of the source image, and the less the distortion of the edge information.	$Q^{AB/F} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} Q^{\overset{A}{F}}(i,j) g_A(i,j) + Q^{\overset{B}{F}} g_B(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} g_A(i,j) + g_{B(i,j)}}$	Higher Value
[35]	FF	Fusion Factor	Is obtained by calculating mutual information between original images and fused image [35].	$FF = MI_{i1,f} + MI_{i2,f}$	Higher value
[35]	FS	Fusion Symmetry	Is applied to compute the degree of symmetry between the data of two photos.	$F_s = abs - \left\lvert \frac{MI_{i1,f}}{MI_{i1,f} + MI_{i2,f}} - 0.5 \right\rvert$	Higher value

Perceptual Quality Metrics					
Ref.	**Metrics**	**Full Name**	**Description**	**Expression**	**Desired Value**
[59]	SSIM	Structural Similarity Index	It is employed to compare the local distributions of pixel intensities of original and fused images. The scale is between -1 and 1. When it equals 1, the two images are close to each other.	$SSIM = \frac{(2\mu I_r I_f + C_1)(2\sigma I_r I_f + C_2)}{(\mu^2 I_r + \mu^2 I_f + C_1)(\sigma^2 I_r + \sigma^2 I_f + C_2)}$	Higher Value
[39]	VIFF	Visual Information Fidelity for Fusion	Relative to the statistical properties of natural scenes, it is a quality index that quantifies image quality in terms of exchange and sharing of information.	$VIFF = \frac{\sum_{i=1}^{M} I(C_i; F_i \mid Z_i)}{\sum_{i=1}^{M} I(C_i; E_i \mid Z_i)}$	Higher value
[63]	UIQI	Universal Image Quality Index	Quantifies the loss of correlation between the luminance and the contrast between the fused result and source images.	$UIQI = \frac{4\mu_x \mu_y \sigma_{xy}}{(\mu_x^2 + \mu_y^2) + (\sigma_x^2 + \sigma_y^2)}$	Higher value
[64]	$Q_{BRISQUE}$	Blind/Reference less Image Spatial Quality Evaluator	The implementation of blind image quality is aimed at predicting the nature of the fused images without the involvement of the input and ground-truth images as references.	Use the Mean Subtracted Contrast Normalized (MSCN) and Generalized Gaussian Distribution (GGD) equations to explain the model.	Lower value
[65]	Q_E	Piella-Heijmans's similarity-based metric	It measures the ability of a fused image to carry meaningful and valuable structural and perceptual features of source images by paying more attention to visually important areas, that is, edges and textures.	$Q_E = \sum_i w_i . [\lambda_i Q(F, A)_i + (1 - \lambda_i) Q(F, B)_i]$	Higher value
[65]	Q_G	Xydeas-Petrovic's gradient-based metric	It is applied to quantify the quality of image fusion based on the quality of how the information of the edge (strength and orientation of the gradient) carried in the source images was preserved in the fused image.	$Q_G = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [Q_G^A(i,j).W^A(i,j) + Q_G^B(i,j).W^B(i,j)]$	Higher value
[54]	SCD	Structure Content Difference	It is used for measuring the structural degradation between the fused and reference image. It measures the degree of loss, relative to the structural content, usually measured by using some gradient property or by the energy content.	$SCD = r(D_1, A) + r(D_2, B)$	Higher value
[65]	Q_{EP}	Edge Information Preservation	Used to measure the edge information preservation of the fusion method by multi-scale evaluation.	$QEP = \frac{\sum_{i=1}^{N} (E_f(i). \max(E_A(i), E_B(i)))}{\sum_{i=1}^{N} \max(E_A(i), E_B(i))}$	Higher value
Statistical and Signal-Based Metrics					
Ref.	**Metrics**	**Full Name**	**Description**	**Expression**	**Desired Value**
[59]	PSNR	Peak Signal-to-Noise Ratio	It is computed by dividing the gray level of the image by the corresponding pixels in both references and fused pictures.	$PSNR = 10 \log_{10} \left[\frac{255^2}{\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2} \right]$	Higher Value
[59]	MSE	Mean Square Error	It is calculated to ensure that the fused and original images are in variation or not.	$MSE = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2$	Lower Value
[59]	RMSE	Root Mean Square Error	It is usually applied to measure the difference between the reference and the fused images with direct computation of variation in pixel values. When the RMSE value is zero, the fused image is near to the original image. A good indicator of the spectral quality of the fused image is RMSE.	$RMSE = \sqrt{\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2}$	Lower Value
[9]	STD / SD	Standard Deviation	It is used to calculate the total contrast of the fused image and serves the purpose of calculating the variation between the data and the average.	$SD = \left[\frac{1}{MN - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} (I(i,j) - \mu I)^2 \right]^{\frac{1}{2}}$	Higher Value
[9]	EN	Entropy	Entropy refers to the contents of the information in the image. It corresponds to the number of points of information found in an image, and it is between 0 and 8.	$E = -\sum_{j=1}^{2^l - 1} p(S_j) \log_2(p(S_j))$	Higher Value
[47]	Q_{CI}	Quality Contrast Index	Is an image fusion assessment term applied to large markers on the increase of contrast in the fused image compared to the reference images. It is a measure of the ability of the fusion process to preserve, or enhance, the contrast between regions of an image, which is of significance to the visual clarity and the diagnostic utility.	$Q_{CI} = \frac{C_F}{\frac{1}{2}(C_A + C_B)}$	Higher value

[9]	SF	Spatial Frequency	Distributes the fused image sharpness, i.e., the change rate of the image gray; the higher the SF is, the higher the image resolution is.	$SF = \sqrt{(RF)^2 + (CF)^2}$	Higher Value
[47]	Q_{AG}	Average Gradient	It is also one of the widely used metrics in image fusion and assessment of image quality in terms of how sharp the image is and how well the details are in it. It is an indicator of the mean rate of change in intensity of neighboring pixels and a sign of the existence of edges and smaller textures.	$Q_{AG} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{(\Delta_{I,x}^2 - \Delta_{I,y}^2)^{\frac{1}{2}}}{2}$	Higher Value
[66]	NC	Normalized Correlation	It is a similar measure that estimates the correlation between two images, usually between the fused and source images. It measures the degree to which the pixel intensity pattern of two images is similar, with corrections made in order to address brightness variations.	$NC = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} A(i,j).F(i,j)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A(i,j)^2}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} F(i,j)^2}}$	Higher Value
[61]	API	Average Pixel Intensity	It is calculated to examine the contrast of the fused photo. The higher API value means the good contrast of the fused photo.	$API = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f(i,j)}{mn}$	Higher value
[47]	Q_{ALI}	Average Light Intensity	Is a simple and significant indicator employed in the process of combining images and upholding images to measure the extent of brightness of an entire image. It gives an estimation of the visual balance of the fused picture.	Typically the same as API, but with luminance scaling in color images	Higher value
[67]	EI	Edge Intensity	Is one of the metrics in image fusion to measure the sharpness and the clarity of the edges of the fused image. It is a measure of transitions between near-neighbor pixel intensities and so may indicate significant properties in a medical image, such as boundaries between anatomical structures.	$EI = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\lvert G(i,j)\rvert$	Higher value
[68]	PC	Phase Congruency	Refer to the standard deviation and the phase congruency of an image, respectively.	$PC(x) = \frac{\sum_n W_n(x)\lfloor A_n(x)\cos(\emptyset_n(x)-\emptyset(x))-T\rfloor}{\sum_n A_n(x)+\epsilon}$	Higher value
Medical and Clinical Metrics					
Ref.	**Metrics**	**Full Name**	**Description**	**Expression**	**Desired Value**
[69]	D. Coe	Dice Coefficient	It is a way to measure the overlap of two binary sets, particularly in image segmentation and medical image fusion. In fusion or segmentation testing, it quantifies the degree to which the fused or predicted image agrees with some ground truth reference.	$Dice = \frac{2.\lvert A \cap B\rvert}{\lvert A\rvert + \lvert B\rvert}$	Higher value
[70]	J.Coe	Jaccard Coefficient	It is a popular measurement that is utilized to assess the similarity and overlap among two sets. It is often applied in image fusion and segmentation used to determine the extent of fidelity between the fused or segmented image and a reference or ground truth.	$Jac = \frac{\lvert K \cap D\rvert}{\lvert K \cup D\rvert}$	Higher value
[71]	ACC	Accuracy	It is a measure of overall performance evaluation, and it is simply a ratio between the number of correctly predicted cases and the total number of cases.	$Acc = \frac{TP + TN}{TP + FN + FN + TN}$	Higher value
[72]	SEN	Sensitivity	Measures the percentage of true positives that are detected properly. In medical imaging, it determines the ability of the model to identify some disease or pathology.	$Sen = \frac{TP}{TP + FN}$	Higher value
[72]	SPE	Specificity	Measures the proportion of negative labels is correctly classified. It explains why the model performs well to avoid a false positive.	$Spec = \frac{TN}{TN + FP}$	Higher value
Optimization and Training Metrics					
Ref.	**Metrics**	**Full Name**	**Description**	**Expression**	**Desired Value**
[42]	L_{grad}	Gradient Loss	Measures the difference in edge patterns (gradients) of the fused image and the source image. It imposes a loss on sharpness, edge, and texture differences, which motivates the fusion network to retain edge information.	$L_{grad} = \sum_{i,j}(\lvert\nabla_x I_f(i,j) - \nabla_x I_s(i,j)\rvert + \lvert\nabla_y I_f(i,j) - \nabla_y I_s(i,j)\rvert)$	Lower value

6.1 Information-theoretic metrics

Information-theoretic metrics assess how much relevant or shared information is retained between the source images and the fused image. These metrics are based on entropy, mutual information, and statistical dependence, and are particularly useful in quantifying the effectiveness of multimodal image fusion in terms of data preservation and correlation.

6.2 Perceptual quality metrics

Perceptual quality metrics evaluate the fused image based on how it would be perceived by human observers. These metrics typically measure structure, texture, sharpness, and overall visual fidelity. Metrics such as SSIM and VIF are widely used to assess whether fusion preserves structural content and perceptual clarity, especially in medical diagnostics.

6.3 Statistical and signal-based metrics

These metrics are based on simple statistics and frequency analysis to measure pixel intensity distributions, contrast, edges, gradients, and noise. They can be used to assess improvements in visual resolution and signal fidelity, especially in resolution improvement tasks.

6.4 Medical and clinical metrics

Medical and clinical metrics are evaluation measures used to assess, which are applied to determine the clinical reliability and diagnostic utility of the fused images. They include classification and segmentation measures such as accuracy, sensitivity, specificity, and overlap-based indices like Dice and Jaccard. These are crucial for validating the performance of fusion algorithms in practical medical applications.

6.5 Optimization and training metrics

These metrics are primarily used during the training of DL-based fusion models. They evaluate how well the network optimizes the objective function, with gradient loss guiding the model to preserve edges and textures or match high-level perceptual features.

No single evaluation metric is sufficient to comprehensively assess multimodal medical image fusion performance. The measurement of information preservation uses information-theoretic metrics, the measurement of visual quality uses perceptual metrics, the measurement of signal fidelity uses statistical metrics, and the measurement of diagnostic usefulness uses clinical metrics. Because individual metrics quantify different image properties, a fusion method may perform favorably according to one criterion while exhibiting weaker performance according to another. Therefore, multiple complementary metrics should be considered to achieve a balanced and reliable evaluation.

In addition to complementary use of evaluation metrics, the reviewed studies also show that network architecture, fusion strategies, and evaluation methodology are closely coupled. CNN-based architectures remain the predominant backbone and are mostly applied in the form of feature-level fusion, which is effective in preserving complementary information from multiple imaging modalities. With the development of newer architectures, such as GAN-based, hybrid, and State Space Model-based architectures, adaptive feature fusion mechanisms, including attention-guided, multi-scale, and cross-modal interaction, are becoming more common to improve feature representation and fusion quality. In all the architecture categories, PSNR and SSIM are the most commonly used evaluation metrics, and recent studies have increasingly relied on information-theoretic, perceptual, and task-based measures to supplement them to get a thorough assessment of the fusion performance.

7. DISCUSSION

7.1 Architecture

Recent advancements in DL have spurred the emergence of a wide range of architectures specifically tailored to multimodal medical image fusion. Such architectures vary in their structural design, supervision methods and computational behavior, yet all portray important spatial and semantic issues in clinical imaging. From traditional convolutional designs to the newer hybrid models, this architectural variety reflects the current active efforts to integrate successfully all four imaging modalities, namely, CT, MRI, PET, and SPECT.

CNNs are still the core of the majority of image fusion systems owing to their demonstrated ability to extract features in a hierarchical manner. For example, M4FNet includes multi-receptive-field modules that allow the representation of local and global features at the anatomical scale [30]. IFCNN uses a hierarchical plan with effective interpolation tools, which maximize fusion with high-resolution inputs [31]. Multi-scale feature extractors trained with CNNs have the advantage of using pretrained models to speed up learning and encode salient spatial features [32], but CNN-VGG19 uses its deep architecture to maximize the semantic consistency of cross-modal features [33]. In paired learning, Siamese CNNs match modalities effectively [34], while the multi-objective differential evolutionary CNN combines fuzzy logic with evolutionary approaches to enhance fusion quality [53].

Several task-specific CNN variants have also been developed. Multi-scale fusion feature models promote consistency of structure with resolution-consistent alignment mechanisms [36], and weighted CNNs by attention promote content in terms of spatial detail retention [37]. Deep CNNs combined with contrast enhancement modules are oriented on salient region visibility, which helps in the interpretation of diagnosis [38]. Residual attention CNNs promote resilience in noisy or complicated settings through residual-attention [39]. Unsupervised CNNs [40] and deep CNNs [41] prove to be reliable even in the absence of labeled training data and this makes them more useful in experiments with scarce ground truth. Generative methods such as MedFusionGAN [42] and dense block-based GANs [43] incorporate adversarial training to enhance visual plausibility, reduce artifacts, and preserve fine anatomical details.

Hybrid architectures represent a significant shift toward more integrated and versatile design paradigms. CNN-transformer hybrids [44] completely combine the locality of convolutional layers with the long-range semantic modeling of transformers, resulting in more context-aware fusion. Laplacian pyramid CNN-transformer models also provide multi-resolution decomposition, which helps preserve high-frequency structural information [45], and CNN-GAN hybrids

[46] attempt to trade off xenosis and cost, which, like other models, apply to near real-time clinical applications. Transfer learning architectures such as VGG19– ResNet-101 [47] use representations learned on the pretrained models to speed up convergence and increase generalization. Similarly, the VGG19 -SNN ensemble [51] combines both the local and global contextual features to improve the visual quality. Multibranch CNNs with semantic perception and attention [52] use semantic refinement by segmentation to maintain structural and contextual similarity in the fused output. Other hybrid models such as multi-layer feature concatenation CNNs [48], CNN-pyramid models [49] and unified CNN-based pipelines [50] tackle various features of fusion quality by incorporating frequency-sensitive encoding, hierarchical fusion phases and end-to-end segmentation-fusion functions.

Beyond CNN-based architecture, unsupervised models such as DBNs are also applicable, especially where the amount of labeled data is limited. Applied together with fuzzy logic, the DBNs may find informative regions before fusion, which improves entropy and structural preservation [35]. The recently introduced FusionMamba framework [54] represents an emerging state-space-based approach based on its state-space definition. Its reduced latency and dynamic feature modeling indicate potential applicability to computationally constrained medical imaging scenarios.

CNN-based architectures are generally suitable for multimodal medical image fusion tasks because they provide a good hierarchical feature extraction capability with moderate computation. Although more complex to train, GAN-based models are better suited to situations where high visual quality and preservation of fine anatomical details are important. For complex clinical applications that involve both local and global features representation, hybrid architectures can be effectively employed but are more costly than the others. Similarly, transformation and feature-based fusion strategies have lower complexity and stable performance, while optimization- and generative-based fusion strategies typically have higher fusion quality, but at the cost of higher computational and training demands.

Beyond architectural and computational considerations, the reviewed studies also demonstrate the increasing clinical relevance of multimodal medical image fusion. Combining structural information and functional information in a single fused image, the integration of complementary modalities such as MRI, CT and PET has been widely used to aid in the analysis of brain tumors, evaluation of Alzheimer's disease, radiotherapy planning and beyond. The results suggest that future DL architectures should be assessed beyond accuracy in fusion quality, but also by their ability to provide accurate diagnosis, treatment planning, and clinical decision-making in clinical practice.

7.2 Challenges

Although multimodal medical image fusion has advanced considerably, DL-based approaches continue to face several important limitations, as summarized in Table 2. These issues include computational effectiveness, data accessibility, training consistency, input inconsistency, and system construction complexity, with each problem being an impediment to clinical integration and practical implementation.

Computational complexity remains a major limitation. Deep, multi-branch, or recursive fusion architecture such as M4FNet [30], CNN-VGG19 [33], DCNN [41], or CNN-Pyramid [49] require substantial computational resources. This difficulty is also increased in hybrid systems such as CNN-transformer [44] and CNN-GAN [46], in which the latency and memory consumption of processing local features along with global features simultaneously or the reverse is very high. Likewise, models with attention modules or segmentation guidance like residual attention CNNs [39] and multibranch CNNs with semantic perception [52] have higher training and inference times and are not scalable in time-sensitive or resource-limiting settings.

Data dependency represents another major limitation. Supervised networks, such as IFCNN [31], weighted feature fusion CNNs [37], and ensemble networks, such as VGG19-SNN [51], are very dependent on large, labeled, and precisely aligned multimodal data. Such data is not usually available in most clinical settings, especially for less common conditions or in underserviced settings. This dependency reduces the generalizability of models, and it can be difficult to apply them across domains. FusionMamba and DBN-based architectures have undergone relatively limited empirical validation [35, 54], which reduces confidence in their robustness across different modality combinations and imaging conditions. Another problem is training instability, especially in GAN-based architecture. Such non-convergent behaviors as mode collapse, vanishing gradients, or unstable loss surfaces can affect MedFusionGAN [42] and Dense Block GANs [43], and thus these architectures are challenging to optimize effectively. Attention models that are sensitive to segmentation [52] are also susceptible to instability where the input is fvnoisy, multi-scale, or low-contrast, and thus it can adversely affect successful convergence and inter-dataset reproducibility.

Input sensitivity and modality variation also pose further problems to the robustness of these systems. Unlike alternative models, such as CNN-based feature extractors [32] and contrast-enhancing CNNs [38], these models may exhibit reduced performance in the presence of misaligned inputs, modality-specific noise, or variations in acquisition parameters. Similarly, CNN-GAN hybrids [46] and unified CNN fusion frameworks [50] can be found to be less adaptable to previously unseen modality combinations or clinical settings, without retraining or fine-tuning. Finally, the optimization load and design rigidity of most advanced frameworks may be a barrier to implementation. CNN -GAN hybrids [46] and fuzzy-enhanced DBNs [35] have multi-stage fusion strategies with manually designed parameters or rule sets and are thus more complex to reproducibly accomplish as well as more time-consuming to integrate. In addition, other architectures like multibranch CNNs [52] and FusionMamba [54] are based on a series of components that depend on each other, which tends to propagate error effects and coordination problems during model training or practical deployment.

7.3 Future directions

Future research should directly address the challenges identified in Section 7.2 by developing lightweight architectures to reduce computational complexity, adopting self-supervised learning to reduce dependence on labeled datasets, improving model interpretability through explainable artificial intelligence (XAI), and enhancing generalization across imaging modalities. These research directions directly address the limitations of current multimodal medical image

fusion techniques.

Reducing computational complexity represents an important research priority. Lightweight backbone architectures such as MobileNet or EfficientNet may provide computationally efficient alternatives to more complex CNN architectures, like VGG19 [33] or multiscale CNNs [36]. These replacements can dramatically decrease inference time and memory consumption and be used in real-time and resource-constrained clinical settings when combined with methods such as model pruning, quantization, and knowledge distillation. Nevertheless, the major technical challenge is to achieve these computational savings without compromising fusion accuracy, structural detail preservation, or diagnostic reliability.

Limited availability of labeled multimodal medical data remains a major constraint, particularly for supervised learning approaches. To mitigate this limitation, semi-supervised and self-supervised learning strategies could be incorporated into models such as IFCNN [31] and CNN-based weighted fusion approaches [37]. Pseudo-labeling, contrastive learning, and cross-modal augmentation are methods that may reduce dependence on large labeled datasets. Additionally, it might be possible to incorporate domain adaptation layers and modality-invariant encoders to enable networks to operate with high reliability in the face of previously unseen or noise-corrupted inputs to enhance reliability in a wide range of clinical settings. One of the outstanding challenges is the ability of these self-supervised and domain-adaptive models to generalize robustly to a wide range of imaging modalities, acquisition protocols, and healthcare institutions.

Complete stabilization of GAN-based training is also of critical importance; various techniques, such as Wasserstein loss functions, spectral normalization, or progressive learning schedules, can be used to address the problem of convergence instability in networks, such as MedFusionGAN [42] and attention-based GANs [52]. Despite these improvements, it is still a research challenge to achieve stable adversarial training while maintaining anatomical fidelity and avoiding image artifacts. Similarly, multi-objective optimization frameworks can also be applied to fuzzy-rule-enhanced DBNs [35] to reduce manual tuning and enhance the overall training stability.

Generalization and interpretability should be improved to enhance clinical translation. Future models should replace static fusion rules with adaptive strategies that dynamically learn fusion weights under different input conditions. For example, learnable token gating or attention distillation could improve the hybrid CNN-transformer model [44] by enabling more effective encoding of global and local features. Despite their success, transformer-based models are still very resource-intensive to compute and memory-constrained, particularly when working with high-resolution multimodal medical images. It is also possible to optimize models, such as FusionMamba [54], to be more suitable for real-world applications by making use of meta-learning, neural architecture search, and hardware-aware optimization. State-space model-based approaches are also a new architecture that needs to be validated in various clinical datasets to ensure robustness and generalizability. Finally, the fusion decisions will be more transparent and clinically reliable through the inclusion of XAI methods like saliency maps, uncertainty estimates, or feature attribution. However, a significant problem is to provide meaningful explanations that accurately reflect the decision-making process for increasingly complex deep-learning models. Further model testing must use varied, standardized data that extends beyond CT, MRI and SPECT so that there is wider applicability across imaging modalities and settings [50, 51].

The scientific contribution of this review lies in its structured synthesis of existing DL architectures, fusion strategies, datasets, and evaluation methodologies for multimodal medical image fusion. The proposed organization enables the systematic identification of architectural limitations and underexplored research areas, including explainable AI, lightweight and domain-adaptive models, and standardized evaluation across clinical imaging modalities. This integrated analysis extends beyond descriptive summarization by identifying methodological gaps and research priorities relevant to the development of clinically applicable multimodal fusion systems.

8. CONCLUSION

This review presents a comprehensive analysis of multimodal medical image fusion based on DL network architecture, fusion strategy, medical image data and evaluation metrics. The results demonstrate that CNN-based methods continue to dominate, while hybrid, Transformer-based, and novel state-space architectures are gradually overcoming the shortcomings of CNN in feature representation and computation efficiency. Several difficulties still exist, however, such as generalization across modality combinations, computational demands, limited multimodal datasets, and standardized clinically relevant evaluation. Furthermore, lightweight and modality-aware architecture, self-supervised learning and explainable AI, standardized evaluation and wider clinical validation should be emphasized in future studies. In summary, this review brings together a unified framework to understand recent methodological advances and future research directions for developing effective and clinically useful multimodal medical image fusion.

REFERENCES

[1] Bhosekar, S., Singh, P., Garg, D., Ravi, V., Diwakar, M. (2025). A review of deep learning-based multi-modal medical image fusion. The Open Bioinformatics Journal, 18(1): e18750362370697. https://doi.org/10.2174/0118750362370697250630063814

[2] Peter, J.D., Fernandes, S.L., Thomaz, C.E., Viriri, S. (2019). Computer Aided Intervention and Diagnostics in Clinical and Medical Images. Springer International Publishing. https://doi.org/10.1007/978-3-030-04061-1

[3] Sandhya, S., Senthil Kumar, M., Karthikeyan, L. (2019). A hybrid fusion of multimodal medical images for the enhancement of visual quality in medical diagnosis. In Lecture Notes in Computational Vision and Biomechanics, pp. 61-70. https://doi.org/10.1007/978-3-030-04061-1_7

[4] Farid, M.S., Mahmood, A., Al-Maadeed, S.A. (2019). Multi-focus image fusion using content adaptive blurring. Information Fusion, 45: 96-112. https://doi.org/10.1016/j.inffus.2018.01.009

[5] Liu, Y., Chen, X., Wang, Z., Wang, Z.J., Ward, R.K.,

Wang, X. (2018). Deep learning for pixel-level image fusion: Recent advances and future prospects. Information Fusion, 42: 158-173. https://doi.org/10.1016/j.inffus.2017.10.007
[6] Ma, J.Y., Yu, W., Liang, P.W., Li, C., Jiang, J.J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48: 11-26. https://doi.org/10.1016/j.inffus.2018.09.004
[7] Li, Y., Zhao, J.L., Lv, Z., Li, J.H. (2021). Medical image fusion method by deep learning. International Journal of Cognitive Computing in Engineering, 2: 21-29. https://doi.org/10.1016/j.ijcce.2020.12.004
[8] Meher, B., Agrawal, S., Panda, R., Abraham, A. (2019). A survey on region based image fusion methods. Information Fusion, 48: 119-132. https://doi.org/10.1016/j.inffus.2018.07.010
[9] Huang, B., Yang, F., Yin, M., Mo, X., Zhong, C. (2020). A review of multimodal medical image fusion techniques. Computational and Mathematical Methods in Medicine, 2020: 1-16. https://doi.org/10.1155/2020/8279342
[10] Dogra, A., Goyal, B., Agrawal, S. (2018). Medical image fusion: A brief introduction. Biomedical Pharmacology Journal, 11(3): 1209-1214. https://doi.org/10.13005/bpj/1482
[11] Tan, W., Tiwari, P., Pandey, H.M., Moreira, C., Jaiswal, A.K. (2020). Multimodal medical image fusion algorithm in the era of big data. Neural Computing and Applications, 37(28): 22995-23015. https://doi.org/10.1007/s00521-020-05173-2
[12] Radu, C., Fisher, P., Mitrea, D., et al. (2020). Integration of real-time image fusion in the robotic-assisted treatment of hepatocellular carcinoma. Biology, 9(11): 397. https://doi.org/10.3390/biology9110397
[13] Hermessi, H., Mourali, O., Zagrouba, E. (2021). Multimodal medical image fusion review: Theoretical background and recent advances. Signal Processing, 183: 108036. https://doi.org/10.1016/j.sigpro.2021.108036
[14] Li, H., Wu, X.J., Kittler, J. (2020). MDLatLRR: A novel decomposition method for infrared and visible image fusion. IEEE Transactions on Image Processing, 29: 4733-4746. https://doi.org/10.1109/tip.2020.2975984
[15] Zhang, B.H., Lu, X.Q., Pei, H.Q., Zhao, Y. (2015). A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform. Infrared Physics & Technology, 73: 286-297. https://doi.org/10.1016/j.infrared.2015.10.004
[16] Zhang, Q., Liu, Y., Blum, R.S., Han, J.G., Tao, D. (2017). Sparse representation based multi-sensor image fusion: A review. arXiv preprint arXiv:1702.03515. https://doi.org/10.48550/ARXIV.1702.03515
[17] Tang, Q., Liang, J., Zhu, F.Q. (2023). A comparative review on multi-modal sensors fusion based on deep learning. Signal Processing, 213: 109165. https://doi.org/10.1016/j.sigpro.2023.109165
[18] Azam, M.A., Khan, K.B., Salahuddin, S., et al. (2022). A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in Biology and Medicine, 144: 105253. https://doi.org/10.1016/j.compbiomed.2022.105253
[19] Zhou, T., Cheng, Q.R., Lu, H.L., Li, Q., Zhang, X.X., Qiu, S. (2023). Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine, 160: 106959. https://doi.org/10.1016/j.compbiomed.2023.106959
[20] Du, J., Li, W.S., Lu, K., Xiao, B. (2016). An overview of multi-modal medical image fusion. Neurocomputing, 215: 3-20: https://doi.org/10.1016/j.neucom.2015.07.160
[21] Tirupal, T., Mohan, B.C., Kumar, S.S. (2021). Multimodal medical image fusion techniques—A review. Current Signal Transduction Therapy, 16(2): 142-163. https://doi.org/10.2174/1574362415666200226103116
[22] Duan, J.W., Xiong, J.Q., Li, Y.H., Ding, W.P. (2024). Deep learning based multimodal biomedical data fusion: An overview and comparative review. Information Fusion, 112: 102536. https://doi.org/10.1016/j.inffus.2024.102536
[23] Luo, F., Zhao, B.J., Fuentes, J., et al. (2025). A review on multi-focus image fusion using deep learning. Neurocomputing, 618: 129125. https://doi.org/10.1016/j.neucom.2024.129125
[24] Nejati, M., Samavi, S., Shirani, S. (2015). Multi-focus image fusion using dictionary-based sparse representation. Information Fusion, 25: 72-84. https://doi.org/10.1016/j.inffus.2014.10.004
[25] Sun, J.G., Han, Q.L., Kou, L., Zhang, L.G., Zhang, K.J., Jin, Z.L. (2018). Multi-focus image fusion algorithm based on Laplacian pyramids. Journal of the Optical Society of America A, 35(3): 480. https://doi.org/10.1364/josaa.35.000480
[26] Agrawal, D., Karar, V. (2019). Bispectral image fusion using multi-resolution transform for enhanced target detection in low ambient light conditions. Indian Journal of Pure & Applied Physics, 57(1): 33.
[27] Rao, B.S.N., Raju, N.V.K., Dhanush, M., Harshith, P.N.S.M., Mehdi, M.J. (2022). MRI and SPECT medical image fusion using wavelet transform. In 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1690-1696. https://doi.org/10.1109/icces54183.2022.9835857
[28] Kalamkar, S. (2023). Multimodal image fusion: A systematic review. Decision Analytics Journal, 9: 100327. https://doi.org/10.1016/j.dajour.2023.100327
[29] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444. https://doi.org/10.1038/nature14539
[30] Ding, Z.S., Li, H.Y., Guo, Y., Zhou, D.M., Liu, Y.Y., Xie, S.D. (2023). M4FNet: Multimodal medical image fusion network via multi-receptive-field and multi-scale feature integration. Computers in Biology and Medicine, 159: 106923. https://doi.org/10.1016/j.compbiomed.2023.106923
[31] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X.L., Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion, 54: 99-118. https://doi.org/10.1016/j.inffus.2019.07.011
[32] Xia, K., Yin, H., Wang, J. (2018). A novel improved deep convolutional neural network model for medical image fusion. Cluster Computing, 22(S1): 1515-1527. https://doi.org/10.1007/s10586-018-2026-1
[33] Elzeki, O.M., Abd Elfattah, M., Salem, H., Hassanien, A.E., Shams, M. (2021). A novel perceptual two layer

image fusion using deep learning for imbalanced COVID-19 dataset. PeerJ Computer Science, 7: e364. https://doi.org/10.7717/peerj-cs.364

[34] Liu, Y., Chen, X., Cheng, J., Peng, H. (2017). A medical image fusion method based on convolutional neural networks. In 2017 20th International Conference on Information Fusion (Fusion), Xi'an, China, pp. 1-7. https://doi.org/10.23919/icif.2017.8009769

[35] Kaur, M., Singh, D. (2019). Fusion of medical images using deep belief networks. Cluster Computing, 23(2): 1439-1453. https://doi.org/10.1007/s10586-019-02999-x

[36] Li, Y., Zhao, J., Lv, Z., Pan, Z. (2021). Multimodal medical supervised image fusion method by CNN. Frontiers in Neuroscience, 15. https://doi.org/10.3389/fnins.2021.638976

[37] Kahol, A., Bhatnagar, G. (2024). Deep learning-based multimodal medical image fusion. In Data Fusion Techniques and Applications for Smart Healthcare, Elsevier, pp. 251-279. https://doi.org/10.1016/b978-0-44-313233-9.00017-5

[38] Bhutto, J.A., Guosong, J., Rahman, Z., Ishfaq, M., Sun, Z., Soomro, T.A. (2024). Feature extraction of multimodal medical image fusion using novel deep learning and contrast enhancement method. Applied Intelligence, 54(7): 5907-5930. https://doi.org/10.1007/s10489-024-05431-z

[39] Li, W.S., Peng, X.X., Fu, J., Wang, G.F., Huang, Y.P., Chao, F.F. (2022). A multiscale double-branch residual attention network for anatomical–functional medical image fusion. Computers in Biology and Medicine, 141: 105005. https://doi.org/10.1016/j.compbiomed.2021.105005

[40] Xu, H., Ma, J. (2021). EMFusion: An unsupervised enhanced medical image fusion network. Information Fusion, 76: 177-186. https://doi.org/10.1016/j.inffus.2021.06.001

[41] Liang, N.N. (2024). Medical image fusion with deep neural networks. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-58665-9

[42] Safari, M., Fatemi, A., Archambault, L. (2023). MedFusionGAN: Multimodal medical image fusion using an unsupervised deep generative adversarial network. BMC Medical Imaging, 23(1): 203. https://doi.org/10.1186/s12880-023-01160-w

[43] Zhao, C., Wang, T., Lei, B. (2020). Medical image fusion method based on dense block and deep convolutional generative adversarial network. Neural Computing and Applications, 33(12): 6595-6610. https://doi.org/10.1007/s00521-020-05421-5

[44] Luo, X., Fu, G.Z., Yang, J.X., Cao, Y.P., Cao, Y.L. (2023). Multi-modal image fusion via deep Laplacian pyramid hybrid network. IEEE Transactions on Circuits and Systems for Video Technology, 33(12): 7354-7369. https://doi.org/10.1109/tcsvt.2023.3281462

[45] Wang, K.P., Zheng, M.Y., Wei, H.Y., Qi, G.Q., Li, Y.Y. (2020). Multi-modality medical image fusion using convolutional neural network and contrast pyramid. Sensors, 20(8): 2169. https://doi.org/10.3390/s20082169

[46] Beatrice, S., Janaki., M.M., Dhivviyanandnam, I. (2025). An automated multimodal medical image fusion framework for Alzheimer detection using deep learning. Journal of Combinatorial Mathematics and Combinatorial Computing, 126, 73-91. https://doi.org/10.61091/jcmcc126-04

[47] Do, O.C., Luong, C.M., Dinh, P.H., Tran, G.S. (2024). An efficient approach to medical image fusion based on optimization and transfer learning with VGG19. Biomedical Signal Processing and Control, 87: 105370. https://doi.org/10.1016/j.bspc.2023.105370

[48] Dinh, P.H. (2025). MIF-BTF-MRN: Medical image fusion based on the bilateral texture filter and transfer learning with the ResNet-101 network. Biomedical Signal Processing and Control, 100: 106976. https://doi.org/10.1016/j.bspc.2024.106976

[49] Liang, X.C., Hu, P.Y., Zhang, L.G., Sun, J.G., Yin, G.S. (2019). MCFNet: Multi-layer concatenation fusion network for medical images fusion. IEEE Sensors Journal, 19(16): 7107-7119. https://doi.org/10.1109/jsen.2019.2913281

[50] Balasubramaniam, S., Chirchi, V., Sivakumar, T.A., Gururama, S.P., Duraimutharasan, N. (2025). Medical image fusion using unified image fusion convolutional neural network. International Journal of Intelligent Systems, 2025(1): 4296751. https://doi.org/10.1155/int/4296751

[51] Allapakam, V., Karuna, Y. (2024). An ensemble deep learning model for medical image fusion with Siamese neural networks and VGG-19. PLoS ONE, 19(10): e0309651. https://doi.org/10.1371/journal.pone.0309651

[52] Lin, C., Chen, Y.J., Feng, S.L., Huang, M.X. (2024). A multibranch and multiscale neural network based on semantic perception for multimodal medical image fusion. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-68183-3

[53] Kaur, M., Singh, D. (2020). Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks. Journal of Ambient Intelligence and Humanized Computing, 12(2): 2483-2493. https://doi.org/10.1007/s12652-020-02386-0

[54] Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z. (2024). FusionMamba: Dynamic feature enhancement for multimodal image fusion with Mamba. Visual Intelligence, 2(1). https://doi.org/10.1007/s44267-024-00072-9

[55] Vakalopoulou, M., Christodoulidis, S., Burgos, N., Colliot, O., Lepetit, V. (2023). Deep learning: Basics and convolutional neural networks (CNNs). Neuromethods, pp. 77-115. https://doi.org/10.1007/978-1-0716-3195-9_3

[56] Sarvamangala, D.R., Kulkarni, R.V. (2022). Convolutional neural networks in medical image understanding: A survey. Evolutionary Intelligence, 15: 1-22. https://doi.org/10.1007/s12065-020-00540-3

[57] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1): 53-65. https://doi.org/10.1109/msp.2017.2765202

[58] Li, Y., El Habib Daho, M., Conze, P.H., et al. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine, 177: 108635. https://doi.org/10.1016/j.compbiomed.2024.108635

[59] Jagalingam, P., Hegde, A.V. (2015). A review of quality metrics for fused image. Aquatic Procedia, 4: 133-142. https://doi.org/10.1016/j.aqpro.2015.02.019

[60] Hossny, M., Nahavandi, S., Creighton, D. (2008). Comments on 'Information measure for performance of

image fusion.' Electronics Letters, 44(18): 1066-1067. https://doi.org/10.1049/el:20081754
[61] Bhutto, J.A., Tian, L.F., Du, Q.L., Sun, Z.Z., Yu, L.B., Soomro, T.A. (2022). An improved infrared and visible image fusion using an adaptive contrast enhancement method and deep learning network with transfer learning. Remote Sensing, 14(4): 939. https://doi.org/10.3390/rs14040939
[62] Kovesi, P. (1999). Image features from phase congruency. Videre: Journal of Computer Vision Research, 1(3): 1-26.
[63] Wang, Z., Bovik, A.C. (2002). A universal image quality index. IEEE Signal Processing Letters, 9(3): 81-84. https://doi.org/10.1109/97.995823
[64] Sun, T., Zhu, X., Pan, J.S., Wen, J., Meng, F. (2015). No-Reference Image Quality Assessment in Spatial Domain. In Proceeding of the Eighth International Conference on Genetic and Evolutionary Computing, Nanchang, China, pp. 381-388. https://doi.org/10.1007/978-3-319-12286-1
[65] Piella, G., Heijmans, H. (2003). A new quality metric for image fusion. In Proceedings 2003 International Conference on Image Processing, Barcelona, Spain, pp. III-173. https://doi.org/10.1109/icip.2003.1247209
[66] Xydeas, C.S., Petrović, V. (2000). Objective image fusion performance measure. Electronics Letters, 36(4): 308-309. https://doi.org/10.1049/el:20000267
[67] Gad, M., Zaki, A., Sabry, Y.M. (2017). Silicon photonic mid-infrared grating coupler based on silicon-on-insulator technology. In 2017 34th National Radio Science Conference (NRSC), Alexandria, Egypt, pp. 400-406. https://doi.org/10.1109/NRSC.2017.7893509
[68] Liu, Z., Forsyth, D.S., Laganière, R. (2008). A feature-based metric for the quantitative evaluation of pixel-level image fusion. Computer Vision and Image Understanding, 109(1): 56-68. https://doi.org/10.1016/j.cviu.2007.04.003
[69] Ma, J., He, Y.T., Li, F.F., Han, L., You, C.Y., Wang, B. (2024). Segment anything in medical images. Nature Communications, 15: 654. https://doi.org/10.1038/s41467-024-44824-z
[70] Rayed, M.E., Islam, S.M.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked, 47: 101504. https://doi.org/10.1016/j.imu.2024.101504
[71] Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In 19th Australian Joint Conference on Artificial Intelligence, pp. 1015-1021. https://doi.org/10.1007/11941439_114
[72] Akobeng, A.K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. Acta Paediatrica, 96(3): 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x