# An Efficient Deep Learning based Multi-Level Feature Extraction Network for Multi-modal Medical Image Fusion

Syed Munawwar
Department of Electronics and Communication Engineering, Jawaharlal Nehru Technological University Anantapur, Andhra Pradesh, India
Department of Electronics and Communication Engineering, Santhiram Engineering College, Nandyal
munawar.ece@srecnandyal.edu.in

Panyam Vuppu Gopi Krishna Rao
Department of Electronics and Communication Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh, India
p.vgopikrishnarao@rgmcet.edu.in

**Abstract:** *Medical image fusion is the process of creating a single image from the information included in several medical images of the same body region taken using various imaging modalities like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Single-Photon Emission Computed Tomography (SPECT) And Positron Emission Tomography (PET). Many deep learning-based techniques for combining medical images have been presented, but creating suitable fusion rules is still challenging. Another issue with single-scale networks is inadequate feature extraction. Therefore, this paper proposes an efficient deep learning-based multi-level feature extraction network for Multi-modal Medical Image Fusion (MMIF). In this research, we propose a new advanced MMIF approach for medical image fusion. The proposed research employs two distinct enhanced Deep Learning (DL) algorithms for low and high-level feature extraction to fully extract and fuse significant and distinctive features from source images. The Improved GoogLeNet (IGoogLeNet) model is used to extract the low-level features, while Modified DenseNet-201 (MDenseNet-201) is used to extract the high-level features. Second, without creating new fusion rules, the proposed unique feature fusion module merely permits the fusion and enhancement of unique features. The Soft Attention (SA) fusion mechanism based on Softmax is used for fusing low-level and high features. Finally, the Modified Resblock module is developed for image reconstruction. For all image pairs, the proposed approach yields average values of 0.7671, 32.84, 19.316, 10.063, 0.8232, 5.3384, and 8.9874 for Edge-based Similarity Measure (QAB/F), Spatial Frequency (SF), Peak Signal-to-Noise Ratio (PSNR), Average Gradient (AG), Structural Similarity Index Measure (SSIM), Mutual Information (MI), and Gradient-based Metric (QG). Compared with the most recently published methods, the experimental findings show that our proposed fusion approach efficiently improves image contrast, brightness, and quality and better retains important information.*

**Keywords:** *Image fusion, deep learning, feature extraction, soft attention, and modified resblock.*

*Received August 21, 2024; accepted February 17, 2025*
*https://doi.org/10.34028/iajit/22/3/2*

## 1. Introduction

Medical image fusion, which combines two or more multi-modal images to increase the accuracy of disease detection, has been extensively studied for clinical medical applications. Therefore, to lessen interference and incorrect judgment when utilizing multiple multi-modal medical images, numerous studies have investigated medical image fusion techniques [12]. Medical areas use a variety of common multi-modal image forms, which are primarily classified into anatomical and functional categories. The two primary anatomical types are Computed Tomography (CT), which provides density information on implants and bones, and Magnetic Resonance Imaging (MRI), which provides a wealth of soft tissue information. Single-Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) images are examples of functional types that offer details on blood flow and organ metabolism [23]. Because every modality has distinct qualities and limits of its own, it is challenging for medical professionals to provide an accurate diagnosis by examining a single modality [9, 10]. For instance, details from multimodal imaging, including those of head, neck, and lung cancers, can surpass PET alone in terms of specificity and sensitivity and must be analyzed to diagnose malignant tumors [16]. When MRI and PET images are combined, one can better identify brain disease by studying the metabolism in small areas of the cortex and obtaining a precise function match. MRI offers a high structural detail resolution [28, 30]. With certain qualities for clinical applications, data from several modalities are combined as part of the medical image fusion process to aid in the analysis of medical issues [2, 25].

There are two categories of image fusion approaches: transform domain and spatial domain [15]. Since spatial domain approaches directly affect image space, they are more susceptible to noise and have a lower capacity to

capture important features; in contrast, Adaptability to these problems is higher for transform domain techniques and has taken the lead in this field [3]. For instance, the three primary processes of the classical method known as Multi-Scale Decomposition (MSD) are multi-scale decomposition, coefficient fusion, and multi-scale reconstruction. Using multi-scale inverse transformation, multi-scale coefficients are used for splitting an image using MSD. The output image is then created by fusing the coefficients using fusion rules [19, 31]. A popular domain image fusion technique for the physiological properties of the human visual system is Sparse Representation (SR) [14]. Transform domain-based existing techniques like the adaptive sparse representation algorithm and SR entail intricate, labour-intensive design challenges that demand a substantial amount of resources and are still comparatively inefficient [33].

Shallow machine learning based on several methods has recently been released [18]. Due to its strong generalization capabilities, image fusion based on Support Vector Machines (SVM) was developed. For example, Wavelet coefficients can be categorized into related features by an adaptive SVM. Image fusion is a common application for neural networks, such as Pulse-Coupled Neural Networks (PCNNs). In this PCNN technique, medical image fusion is achieved by combining PCNN with the conventional Non-Subsampled Contourlet Transform (NSCT) [5]. However, these techniques have several mathematical problems, including enormous parameter sets and non-linearity. In the transform domain, PCNN must also be used in conjunction with other transform techniques. Image fusion has also been utilized with reinforcement learning.

Additionally, after advancements in deep learning, several widely used Deep Learning (DL) networks, such as Convolutional Neural Networks (CNNs), Visual Geometry Group Networks (VGGs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) were presented for image fusion [11]. For instance, CNN is used by the image fusion framework Improved Fusion Convolutional Neural Network (IFCNN) to extract prominent features [17]. To merge two exposure images, a novel, prior-aware Generative Adversarial Network (GAN) is also created. To reconstruct the image, it comprises a detailed guided decoder, and the input image's semantics are encoded using a content-prior-guided encoder [13]. To guarantee that more detailed information is effectively preserved in the final fused image, deep learning techniques offer powerful feature extraction and simple implementation [1]. Consequently, deep learning feature extraction has been used extensively in medical image fusion, offering good performance in obtaining both high-level and low-level features [32]. Most non-end-to-end networks require complex artificial fusion architecture. While fusion rules are not necessary for end-to-end networks, Final fusion and network training are made more difficult by the lack of ground truth [34].

To arrive at proper performance evaluation measures for the objective findings of medical image fusion to be successful, input image details and important data must be correctly preserved. To address these issues, we, therefore, propose a novel and cutting-edge method for Multi-modal Medical Image Fusion (MMIF) that combines deep learning-based feature extraction models for medical image fusion without the need for manually created fusion rules. The proposed approach of separately extracting and fusing low-level and high-level features is a powerful technique in image fusion. It makes use of the advantages of both feature sets to create excellent fused images that are contextually and technically rich.

The following are some of the important contributions of the proposed research:

- A novel and efficient model for MMIF is proposed in this research, it successfully combines multi-modal medical images using deep learning models for feature extraction, a Soft Attention (SA) fusion rule based on Softmax for feature fusion and a Modified Resblock module for image reconstruction.
- Initially, the pre-processing techniques are applied to enhance the quality of the image such as the undesired noise in the images being removed using median filtering and the image contrast is enhanced using the Histogram Equalization (HE) technique.
- After pre-processing, the low-level features are extracted using the Improved GoogLeNet (IGoogLeNet) model while the high-level features are extracted using the Modified DenseNet-201 (MDenseNet-201) model.
- Then for image fusion, the SA fusion mechanism based on Softmax is applied to fuse the high-level and low-level features. The fused high-level and low-level features can be combined to form the final fused image using the Modified Resblock module in the image reconstruction phase.
- We evaluate the proposed method on multiple multi-modal medical image pairs (Magnetic Resonance Imaging-Computed Tomography (MRI-CT), Magnetic Resonance Imaging-Positron Emission Tomography (MRI-PET), and Magnetic Resonance Imaging-Single-Photon Emission Computed Tomography (MRI-SPECT)) using different performance measures and show improved fusion performance in comparison to the nine most advanced fusion techniques.

The remaining parts of this research are organized as follows, section 2 displays an overview of related works on medical image fusion. Details of the proposed approach structure are given in section 3. The experimental results and performance evaluation are made in section 4. Lastly, the paper is summarized and concluded in section 5.

## 2. Related Prior Works

Different MMIF methods were provided by researchers. These methods are investigated and yield accurate outcomes. We summarize and discuss a few of this research in this section.

Using various imaging modalities, a medical image fusion technique was created by Sinha *et al*. [21] in the Non-Subsampled Shearlet Transform (NSST) domain for combining a grayscale image with the corresponding pseudo-color image acquired. To identify the corresponding sub-images, the grayscale image is first decomposed using the NSST. The Maximum Regional Energy (MRE)-based rule and Prewitt operator fuse the low-pass sub-images. The fused high-pass sub-images are obtained using the developed Improved Dual-Channel Pulse-Coupled Neural Network (IDPCNN). In the combined sub-images, the inverse NSST is applied to create the final fused image.

The Principal Component Analysis Network (PCANET) developed by Ghandour *et al*. [6] is a reasonably simple deep-learning model for extracting medical image features. A Principal Component Analysis (PCA) filter is used for feature extraction. For medical image fusion, a useful feature space is developed by this work using PCANET and the nuclear norm. Specifically, the retrieved PCANET properties can function like a CNN. Finally, the Final Decision Map (FDM) is evaluated using a Focus Score Map (FSM). Using FDM, the fused image is created through the combination of the two input medical images.

For medical image fusion, a unified AI-Generated Content (AIGC) system called Cross-Modal Interactive Network (CMINet) was introduced by Song *et al*. [22]. CMINet combines an interactive CNN with a recursive transformer. Within modalities, the extended spatial and temporal dependencies are captured by designing a recursive transformer. Across modalities, the local features are extracted and fused using the interactive CNN. With extensive functional and structural details, the developed approach can produce fused images by making use of cross-modality interaction learning. Furthermore, the recursive network's architecture is designed to minimize the number of parameters, which may be advantageous for implementation on devices with limited resources.

For unsupervised multi-modal medical image fusion, an adaptive cross-modal fusion technique was studied by Xie *et al*. [29]. Specifically, the cross multi-axis attention mechanism-based lightweight cross Transformer is developed in this work. The multi-modal data's local and global interactions are extracted and integrated using the cross-window and cross-grid attention techniques in this paper. A spatial adaptation fusion module provides additional guidance to the cross Transformer, enabling the model to concentrate on the most pertinent data. For feature extraction, a unique feature extraction module is built that integrates many gradient residual dense convolutional and Transformer layers.

A lightweight residual fusion network, a real-time MMIF technique, was presented by He *et al*. [7]. Initially, a three-branch feature extraction architecture is created. The texture and brightness data are fully extracted using the two separate branches. The texture and brightness data can be better preserved by dynamically fusing distinct modalities at a shallow level due to the fusion branch. Moreover, the model's traditional residual convolution is intended to be replaced by a lightweight residual unit for image fusion.

A novel MMIF technique utilizing the Parameter Adaptive-Pulse Coupled Neural Network (PA-PCNN) and Non-Subsampled Contourlet Transform (NSCT) was presented by Ibrahim *et al*. [8]. These images are divided into high- and low-frequency bands by the NSCT. The bands are combined using PA-PCNN. The NSCT approach's inverse was employed to create the fused image. This method's flaw is that the quality metrics aren't appropriate for combining SPECT, MR-T1, and MR-T2 image modalities. The following Table 1 provides an overview of these studies.

Table 1. Literature survey.

| Related works | Year | Approach | Pros | Cons |
|---|---|---|---|---|
| Sinha *et al*. [21] | 2024 | IDPCNN | High performance and low computational time | It is computationally intensive and time-consuming |
| Ghandour *et al*. [6] | 2024 | PCANET | Improves results while requiring less computing time | This technique still adds unnecessary noise to the medical images before fusing them into one final image. Absence of processing steps for pre-and post-processing. |
| Song *et al*. [22] | 2024 | CMINet | Rich structural and functional information are produced in fused images. | Compared to some deep learning algorithms, this method takes longer to execute overall. |
| Xie *et al*. [29] | 2024 | Cross multi-axis attention method based on a lightweight cross Transformer | This model produces a better fusion effect. | (1) Complex architecture<br>(2) Computationally intensive |
| He *et al*. [7] | 2024 | LRFNet | The recommended fusion strategy has been found to outperform the competition. | The drawback of this approach is that combining CT, MRI, and SPECT imaging modalities is not appropriate for quality assessments. |
| Ibrahim *et al*. [8] | 2024 | PA-PCNN | Experiments conducted both qualitatively and quantitatively show that the suggested method is better than other fusion methods. | It takes a lot of time and computing power. |

### 2.1. Problem Statement

An essential component of image-based disease diagnosis is the fusion of medical images. Many approaches for the fusion of medical images have been proposed recently. However, the images produced by

the existing fusion methods frequently have drawbacks, such as poor image quality, a loss of important information, image distortion, decreased brightness and contrast, a loss of edge information, a limited capacity to preserve details, and an enormous requirement for training data for deep learning. In this research, we provide a novel approach to address the previously noted difficulties in the process of medical image fusion. A new advanced MMIF method based on the efficient multi-level feature extraction models, SA fusion mechanism based on Softmax, and Modified Resblock module-based image reconstruction is proposed.

## 3. Proposed Methodology

The primary emphasis of our research is the integration of MRI and PET/CT/SPECT images. Preprocessing must be done before combining SPECT/PET and MRI data since PET and SPECT images are color. In the fusion network, the MRI and CT images are supplied directly. From Red Green Blue (RGB) to Luminance, Chrominance-Blue, Chrominance-Red YCbCr space, the primary purpose of pre-processing is to convert color images. Subsequently, the input image is identified as the *Y* component. Next, by fusing two input images, a fused image *Y Fused* is produced utilizing the proposed MMIF network model. Lastly, the final fused image is produced by transforming the *Cb*, *Cr*, and *Y Fused* components into the Red, Green and Blue (RGB) color space. Figure 1 depicts the proposed MMIF network model's organizational structure.

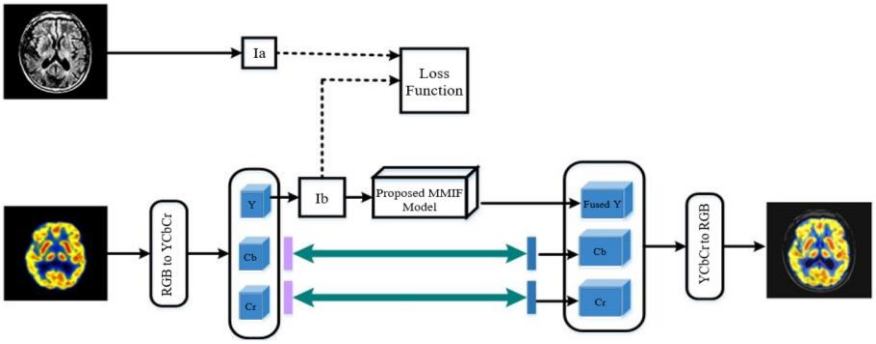



Figure 1. An overview of the proposed MMIF fusion network.

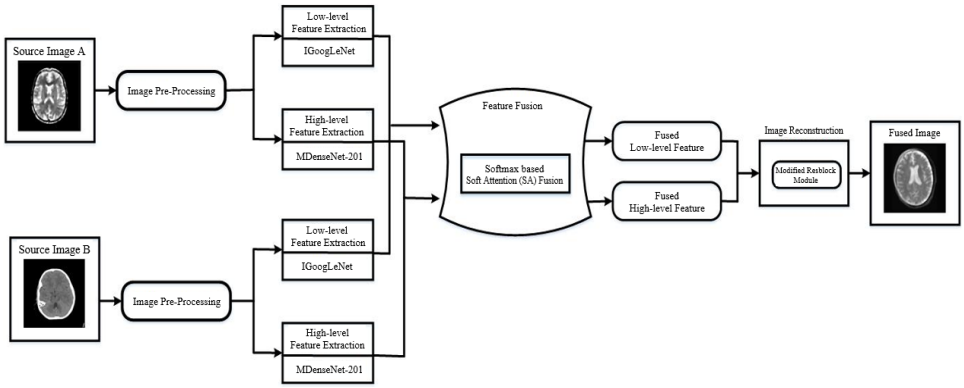



Figure 2. The overall structure of the proposed MMIF fusion approach.

The proposed MMIF network model's intricate network architecture is revealed in Figure 2, which is divided into four primary sections: image pre-processing, feature extraction, feature fusion, and image reconstruction. filtering (median filter) and contrast enhancement HE are two methods used in the pre-processing step to improve the quality of the input images. It has two branches for the feature extraction phase. The IGoogLeNet is included in the upper branch, and MDenseNet-201 is included in the lower branch. Images' low-level features are extracted by the IGoogLeNet model, and their high-level features are extracted by MDenseNet-201. After that, feature fusion is accomplished using attention-based fusion rules, and low- and high-level feature fusion is accomplished using a soft attention fusion rule based on Softmax, the structural and functional details of the fused images are effectively balanced using the proposed fusion mechanism. Using the combined features, this fused image is produced using an image reconstruction module. The final image can be created by combining fused high-level and low-level features, which is accomplished through the usage of the Modified Resblock module during image reconstruction. The proposed model undergoes training and validation using the Whole Brain Atlas dataset. The proposed research's findings are assessed based on impartial evaluation standards. Nine popular cutting-edge methods are

compared with the proposed method to verify its validity and accuracy. Algorithm (1) shows the step-wise algorithm for the proposed fusion approach.

*Algorithm 1. The stepwise algorithm of the proposed fusion approach:*

*Input: MRI and CT image pairs.*
*Output: Fused image with enhanced quality and performance measures.*
*Step 1: Pre-processing*
- *Noise Removal: Apply median filtering to remove noise and preserve edges in the MRI and CT images.*
- *Contrast Enhancement: Use histogram equalization to improve the visual contrast of the input images.*

*Step 2: Feature Extraction*
- *Low-Level Feature Extraction: Extract fine-grained features such as edges and textures using the IGoogLeNet model.*
- *High-Level Feature Extraction: Extract semantic and contextual features using the Modified DenseNet-201 model.*

*Step 3: Feature Fusion*
- *Combine low-level and high-level features.*
- *Use the Soft Attention (SA) mechanism based on Softmax to dynamically assign weights and emphasize important features during fusion.*

*Step 4: Image Reconstruction*
- *Develop a Modified Resblock module to reconstruct the fused image from the fused features.*
- *Ensure that both fine details and semantic information are preserved in the reconstructed image.*

*Step 5: Performance Evaluation*
*Analyze the quality of the fused image using the following performance measures:*
- *QAB/F (Fusion Quality Index)*
- *SF (Spatial Frequency)*
- *PSNR (Peak Signal-to-Noise Ratio)*
- *AG (Average Gradient)*
- *SSIM (Structural Similarity Index Measure)*
- *MI (Mutual Information)*
- *QG (Gradient-Based Quality Metric)*

## 3.1. Database Description

The Whole Brain Atlas Harvard medical dataset, which is a publicly accessible dataset (https://www.med.harvard.edu/aanlib/), was used to train the proposed MMIF approach. Since every image in the collection has already been co-registered, no registration is necessary. 256x256 is the default image size. In the present work, two multimodal images are used as the source image for image fusion. A range of imaging modalities are included in the dataset; for experiment analysis, we have used three sets of pair images, including MRI-CT, MRI-PET, and MRI-SPECT.

## 3.2. Dataset Splitting

To verify the effectiveness of the proposed research, we randomly divide the dataset into training and testing. Tables 3 and 4 illustrate the dataset distribution used for training and testing the proposed network across multiple imaging modalities, including MRI-CT, MRI-PET, and MRI-SPECT.

Algorithm (1) provides the image distribution for network training, covering various disease categories such as metastatic adenocarcinoma, Alzheimer's disease, mild Alzheimer's, glioma, and meningioma. Notably, MRI-SPECT contributes 45 images each for Alzheimer's, mild Alzheimer's, and glioma. In comparison, MRI-CT is used for meningioma with 60 images, summing up to 90 images for MRI-SPECT and 60 images for MRI-CT.

Table 2 presents the dataset distribution for network testing, focusing on Alzheimer's disease, acute stroke, and sub-acute stroke. MRI-CT and MRI-SPECT are used for testing acute stroke and sub-acute stroke cases, respectively, each containing 30 images. Additionally, MRI-PET is utilized for Alzheimer's disease testing with 30 images. Each modality is allocated an equal number of images for evaluation, ensuring a balanced assessment of the network's generalization performance.

Table 2. Image distribution for network training.

| Modalities | Metastatic adenocarcinoma | Alzheimer's | Mild Alzheimer's | Glioma | Meningioma | Total |
|---|---|---|---|---|---|---|
| MRI-CT | | | | | 60 | 60 |
| MRI-PET | | | 45 | 45 | | 90 |
| MRI-SPECT | 45 | 45 | | | | 90 |

Table 3. Image distribution for network testing.

| Modalities | Alzheimer's | Acute stroke | Sub-acute stroke | Total |
|---|---|---|---|---|
| MRI-CT | | 30 | | 30 |
| MRI-PET | 30 | | | 30 |
| MRI-SPECT | | | 30 | 30 |

## 3.3. Image Pre-Processing

Initially, the pre-processing techniques are applied to the original images to enhance the final fused images' visual quality. In the pre-processing stage, we performed noise removal based on Median Filtering and contrast enhancement based on HE.

### 3.3.1. Noise Removal

In the proposed approach, the median filtering technique is used to eliminate undesirable noise from the input images. The median filter is a nonlinear spatial filter. The "speckle" noise in the image could be eliminated by applying a median filter. Because of its de-noising ability and computational sufficiency, it is commonly used in encouraging noise removal approaches. Typically, it's used to minimize noise and smooth out images without causing edge blur. The median filter is one popular non-linear filter used to eliminate Salt and Pepper noise. In the input window, the median brightness value is used to determine the output pixel, which is dependent on the window being moved over the image.

### 3.3.2. Contrast Enhancement

Contrast enhancement enhances the image quality for easier interpretation and makes it easier to retrieve the information it contains. HE, the most popular technique for contrast enhancement, frequently allows for the increase of contrast on image details. Making the image's grey-level histogram as flat as possible is the goal of this change. The following steps explain how to get a consistent dispersion following a point transition:

Preservation of Average Brightness as the basis for Equalization Depending on whether contrast enhancement is applied before or after the restoration procedure, the goal is to improve the image quality while maintaining the average brightness.

The average of an image X, such that $X_m \in \{X_0, X_1,...,X_{L-1}\}$, is indicated by the letter $X_m$. While maintaining the mean brightness, the image X is divided into two sub-images $X_L$ and $X_U$ during the process of HE,

$$X = X_L \cup X_U \quad (1)$$

Following the individual equalization of the two sub-images, the resulting equalized image is composed of the original image's average brightness.

## 3.4. Feature Extraction

In image processing, where feature extraction and representation are crucial steps, deep learning techniques have been applied extensively. Due to the great precision and rich semantic information that different scale features may represent, multi-scale techniques are frequently employed to handle features. Numerous well-known and traditional techniques demonstrated that the introduction of multi-scale feature extraction can lead to superior outcomes.

To improve the quality of images used in clinical diagnosis, medical image fusion combines complementary data with multi-modal images. Typically, there are three steps in this process: image reconstruction, feature fusion, and feature extraction. Nevertheless, the majority of image fusion techniques are unable to adequately adjust for cross-modal interaction when attempting to extract shared and specific data from various modal images, which results in imperfect feature extraction and fusion. Furthermore, most existing approaches' multilevel feature interaction is inadequate, which results in improper usage of fused information under various receptive fields. To overcome these problems, we employ distinct deep learning models to extract the high-level and low-level features, utilizing the advantages of each feature type to create high-quality fused images that are contextually and technically rich. High-level features record semantic information, while low-level features record specific details. The overall fused image quality can be improved by combining both images. The low-level features in the proposed research are extracted using the IGoogLeNet model, and the high-level features are extracted using the MDenseNet-201 model. The fusion process can be more resilient to noise and fluctuations in the input images by using features from many levels.

### 3.4.1. Low-Level Feature Extraction

The low-level features including frequency domain features, intensity histogram features, shape features, spatial characteristics, texture features, and intensity information are extracted using the IGoogLeNet model. Google researchers created the CNN-based architecture known as GoogLeNet. The model demonstrated its strength as the victor of the ImageNet 2014 competition.

The GoogLeNet architecture's primary goal is to attain great accuracy at a low computational cost. The split, transform, and merge concepts used in the inception architecture, which gave rise to CNN's inception block, are the foundation upon which the GoogLeNet model is built. This design combines multi-scale convolutional transformations. Figure 3 shows an overview of the inception block. Other deep learning architectures have set convolution sizes for each layer; this is not the case with the inception module. The 3×3 max pooling and the 1×1, 3×3, and 5×5 convolutions all function in parallel at the input of the inception module, and their combined output is stacked to produce the final result. Large convolutional filters are a computationally and memory-intensive feature of the original GoogLeNet model. Deep network training becomes difficult as a result, particularly on systems with restricted resources. Over-fitting is a common problem with large convolutional filters with multiple parameters. To navigate around these challenges, we improve the GoogLeNet model's structure.

Small blocks are used in place of conventional convolutional layers in the GoogLeNet model. At various scales, the spatial information of the images including both fine and coarse grain levels are captured by using condensed filters of various sizes, such as 1×1, 3×3, and 5×5 in these blocks. Alongside 3×3 max-pooling layers, the GoogLeNet model has numerous convolutions with 5×5 filters, 3×3 filters, and 1×1 filters arranged, as depicted in Figure 3.

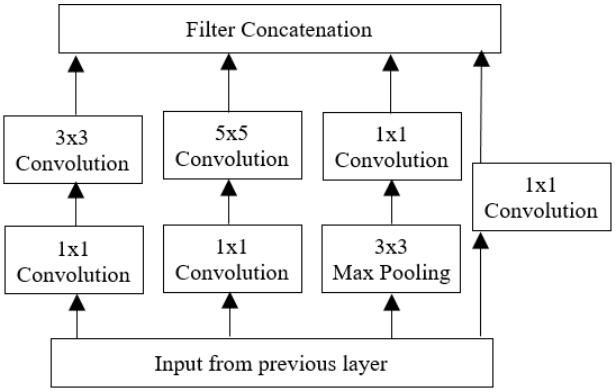



Figure 3. The overview of the inception block.

Before using large-size kernels, the GoogLeNet model adds a bottleneck layer of 1x1 convolutional filters to control the calculations. The architecture's number of parameters (weights and biases) is reduced by using 1×1 convolution. Additionally, it eliminates duplicate information and lowers costs by removing irrelevant feature maps by using sparse connections. Additionally, it eliminates duplicate information and lowers costs by removing irrelevant feature maps by using sparse connections. The number of parameters is significantly reduced as a result of these parameter-tuning processes.

Table 4. The model architecture of IGoogLeNet used in this research.

| Layer | Patch size/ stride | Depth | Pool Proj | Output size |
|---|---|---|---|---|
| Conv 1 | 7x7/2 | 1 | | 112×112×64 |
| MaxPool 1 | 3x3/2 | 0 | | 56×56×64 |
| Conv2 | 3x3/1 | 2 | | 56×56×192 |
| Max pool 2 | 3x3/2 | 0 | | 28×28 ×192 |
| Inception-3a | | 2 | 32 | 28×28×256 |
| Inception-3b | | 2 | 64 | 28×28×480 |
| Max pool3 | 3x3/2 | 0 | | 14×14×480 |
| Inception-4a | | 2 | 64 | 14×14×512 |
| Inception-4b | | 2 | 64 | 14×14×512 |
| Inception-4c | | 2 | 64 | 14×14×512 |
| Inception-4d | | 2 | 64 | 14×14×528 |
| Inception-4e | | 2 | 128 | 14×14×832 |
| Max pool4 | 3x3/2 | 0 | | 7×7×832 |
| Inception-5a | | 2 | 128 | 7×7×832 |
| Inception-5b | | 2 | 128 | 7×7×1024 |

As shown in Table 4, we created the IGoogLeNet model's structure specifically for low-level feature extraction in this research. The IGoogLeNet model's layer-by-layer design components are displayed in this table. The different convolution filters employed in the inception module are 1×1, 3×3, and 5×5. The reduction layer's ×1 filters utilized before related convolution layers are represented by the 3×3 reduce and 5×5 reduce signs. The "pool projection" column, is also referred to as "Pool proj". In the projection layer, it indicates how many 1×1 filters are present following the inherent maximum pooling. The term "max pool" refers to the greatest quantity of pooling layers. These max-pooling layers are designed to down-sample the input as it is transmitted throughout the network. Repaired linear Units (ReLUs) are the activation functions of all the convolution, reduction, and projection layers in this architecture. Without pooling, there are 22 layers in this architecture (or 27 layers if pooling is taken into account). The output of the last inception module yields the extracted low-level features (1x1 convolution layer).

### 3.4.2. High-Level Feature Extraction

Using the MDenseNet-201 model, more semantic information such as shapes, objects, and patterns are extracted. DenseNet-201 presents a dense connection where feature maps from all earlier layers are sent to each layer through a dense block. Reusing features is encouraged by this connectivity topology, which also enhances information flow throughout the network. DenseNets are made up of dense blocks, each of which has a constant component size but varying filters in some cases. The downward sampling is regulated by DenseNet Transition Networks using Batch Normalizing layers (BN), 1×1 convolution, and 2×2 pooling. We evaluate the various deep learning models, select the one that works best, and change it to satisfy our needs. With the addition of specific layers, we change the DenseNet201 model's structure to improve the original DenseNet201 model. Reduced trainable parameters resulted from this, which aided in lowering execution speed and computing complexity.

With one Max-Pooling Layer (MPL) pool size of two, one dense layer, one drop-out layer, and one flatten layer, the original DenseNet201 model was enhanced. Ultimately, this modified DenseNet201 (MDenseNet201) model was used for training. In every map of features, one of MPL's most important tasks was figuring out how much value was contained. When using a pooling process, feature maps and filters are similar. The pooling process is substantially quicker than feature maps. This model demonstrates how employing two MPL layers reduces the size of each feature map. This is a result of its usage of a two-pool size. Figure 4 displays the MDenseNet201 structure.

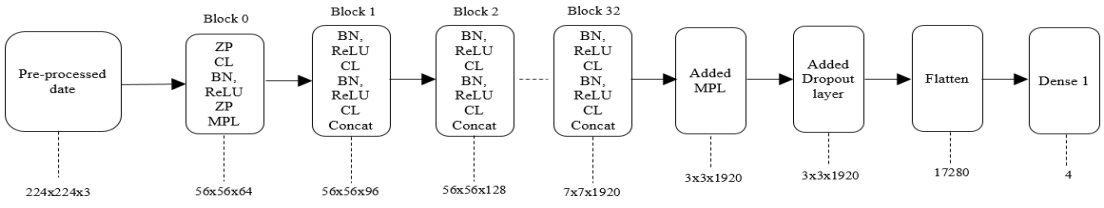



Figure 4. A schematic representation of the MDenseNet-201 model.

In the DenseNet model, the dense block output is transformed into a one-dimensional vector as the primary purpose of the flatten layer, which may be feature maps. The MDensNet201 architecture is made up of 3x3 parts or blocks. Block 0 has 1 has several layers including the convolution, BN, MPL, ReLu, and Zero Padding (ZP). The dense layers were used to categorize the acquired features. The major advantage of our proposed MDenseNet201 has a small number of parameters compared to the original DenseNet201 model.

## 3.5. Feature Fusion

Several image fusion rules can be used to conduct feature fusion. The attention-based fusion method is one

of them. In existing papers, a Motion-Attentive Transition Network (MATNet) was used in the SA fusion for the segmentation process. An Attention Transition (AT) unit and a soft attention unit make up the two units that make up the Motion-Attentive Transition (MAT) module. Appearance learning is promoted by transmitting the attentive motion features by the AT unit. Focusing on important input regions is made easier with the help of the SA unit. Both the high-level and low-level features of $I_1$ and $I_2$ are fused using a Softmax-based soft attention fusion module.

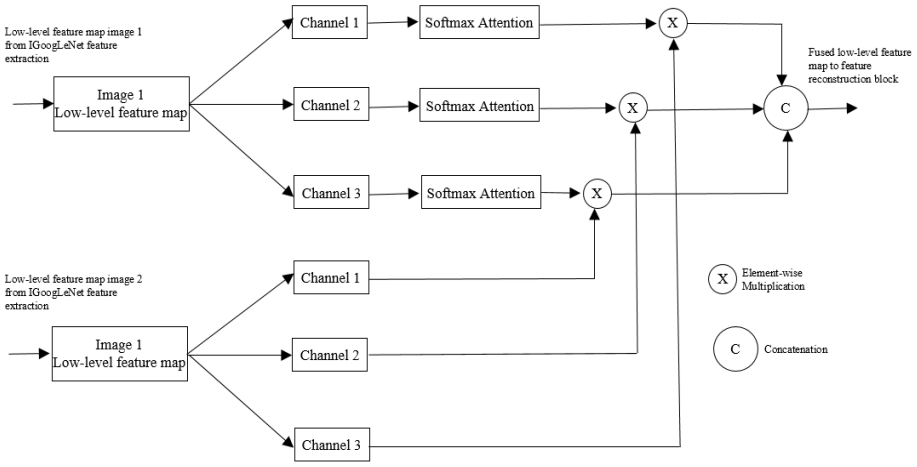



Figure 5. The soft attention fusion block structure.

The SA Fusion module, a small part of the MAT block, is used in this instance to fuse the most information possible into a single image. Figure 5 shows the soft attention fusion module's structure. This figure shows the fusion process for low-level features; high-level features are fused similarly. Initially, the feature maps $I_1$ and $I_2$, which were acquired during the earlier stage of feature extraction, are divided into three channels each. Then, as shown in the bottom (2), the attention map is produced by applying the Softmax function over the channels of the $I_1$ feature map,

$$\begin{matrix} Soft\max \quad attention: (A_{am}) = \\ soft\,max \quad (I_1 featuremapchannels[i]) \end{matrix} \tag{2}$$

Where the channel's index number is depicted as 'i', ranging from 0 to 2.

To improve attention weights-based features, the attention map $A_{am}$ is used, which was produced from the soft-max operation and then the relevant channel of the $I_2$ feature map is multiplied element-wise with it,

$$\begin{matrix} Enhancedfeatures: E_f = \\ (A_{am}) \otimes (I_2 featuremapchannels[i]) \end{matrix} \tag{3}$$

Where the corresponding channel I's attention-aware feature map is represented as, $E_f$ and the element-wise multiplication operation is denoted as $\otimes$, which is then fed into the feature extraction layer to create the fused feature map. After that along the channel direction, each channel's enhanced features are ultimately concatenated. At last, the low-level and high-level fused feature maps have been acquired.

## 3.6. Image Reconstruction

Feeding the fused feature maps into a Modified Resblock module, which has three convolutional layers including 3×3, 3×3, and 1×1. The final resulting fused image is obtained from the third convolutional layer after the feature channels of the first and second layers are combined. The fused low-level feature maps are sent to the first 3x3 convolutional layer, and the fused high-level feature maps are sent to the second 3x3 convolutional layer. The 1×1 convolutional layer combines these feature maps to form the final reconstructed image. The input for the first and second layers of convolution is 64 channels. The output channels of the three convolutional layers are 64, 64, and 32, as can be seen above. Figure 6 displays the image reconstruction module.

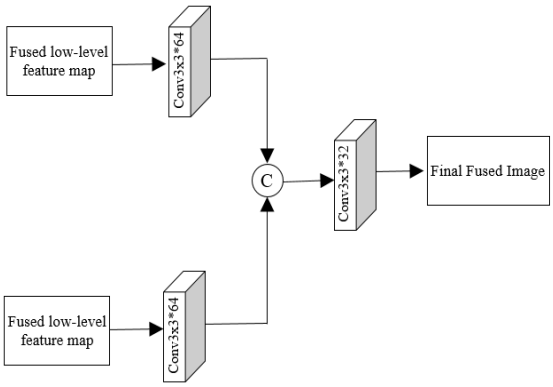



Figure 6. Image reconstruction module.

## 3.7. Loss Functions

In addition to learning the image reconstruction at the pixel level, the network also extracts the image's structure and gradient information. Our foundation is structural similarity, which is represented by the value of SSIM, which ranges from -1 to 1. The closer the value is to 1, the more similar the two are and the higher the quality of the fusion. The image's structural information

is effectively extracted in the proposed approach using SSIM loss, and it is defined as:

$$L_{SSIM} = 1 - SSIM(x, y) \tag{4}$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

Where the input image is represented by $x$, while the reconstructed image is represented by $y$. The variance of x and y is described as $\sigma_x^2$ and $\sigma_y^2$. The mean of $x$ and $y$ is described as $\mu_x$ and $\mu_y$. The covariance of $x$ and $y$ is described as $\sigma_{xy}$. In cases where $\mu_x^2 + \mu_y^2$ and $\sigma_x^2 + \sigma_y^2$ are extremely near to zero, the unstable outcomes are prevented by employing the $C_1$ and $C_2$ variables.

## 4. Result and Discussion

This section discusses the proposed approach's effectiveness and performance. The proposed approach's fusion results are also simulated. Discussions are held over each source image utilized in this research. This part is divided into five subsections: evaluation measures, experimental setups and parameter settings, and performance evaluation of the proposed MMIF approach with a comparison of recently published fusion approaches.

### 4.1. Experimental and Parameter Settings

The proposed research's performance is examined by dividing the dataset into training and testing groups at random. For the analysis of experiments, Python software is employed. The investigation was conducted using a computer equipped with 16 GB of main memory, an NVIDIA GeForce RTX 3060 Ti graphics card, and an Intel Core i5-12400F CPU. During the training phase, 32 is the batch size, 0.0001 is the learning rate, and the number of epochs is 100. The parameters of the proposed network models are optimized using the Adam optimizer.

### 4.2. Evaluation Measures

To precisely and statistically analyze the performance of the proposed approach and assess the fused medical image, seven representative evaluation measures are chosen in the medical image fusion field. The evaluation measures are MI, $Q_G$, SSIM, PSNR, AG, SF, and $Q_{AB/F}$.

$Q_{AB/F}$: It is a comprehensive evaluation of the image fusion quality by measuring the extent to which important textural features and edge information are conserved in the fused image.

- **SF**: By monitoring changes in intensity at various spatial frequencies, SF evaluates an image's texture and amount of detail by successfully differentiating between smoother areas and finer details.
- **PSNR**: A metric indicating how faithfully the original images are matched in a fused image is provided by PSNR
- **AG**: AG measures an image's total intensity change.
- **SSIM**: By considering structure, contrast, and luminance, the similarity between two images is evaluated by using the SSIM measure.
- **$Q_G$**: The source and fused image's gradient information is compared for analyzing the quality of fusion is called $Q_G$. Because the image's edges and textures are represented by the gradient information, $Q_G$ can determine whether the key details and structures of the original image are retained during image fusion.
- **MI**: Between the source and fused images, the shared information content is analyzed using the MI metric. It can be applied to evaluate the image fusion's quality and make sure that the fused image's final representation accurately transmits and incorporates the relevant information.

### 4.3. Performance Evaluation

The Harvard medical dataset, which is openly accessible, is used to assess the proposed method. The efficacy of the proposed model is demonstrated using three different sets of medical images, one pair each of MRI-CT, MRI-PET, and MRI-SPECT.

This technique ensures that an efficient fusion regulation is created, allowing a substantial portion of the information from the input images to be preserved in the fused image. We use the most cutting-edge methods currently available for multi-modal medical image analysis to present both qualitative and quantitative evaluations. This new method creates fused images that are rich in texture features, noise-free, and able to identify the exact tumor region.

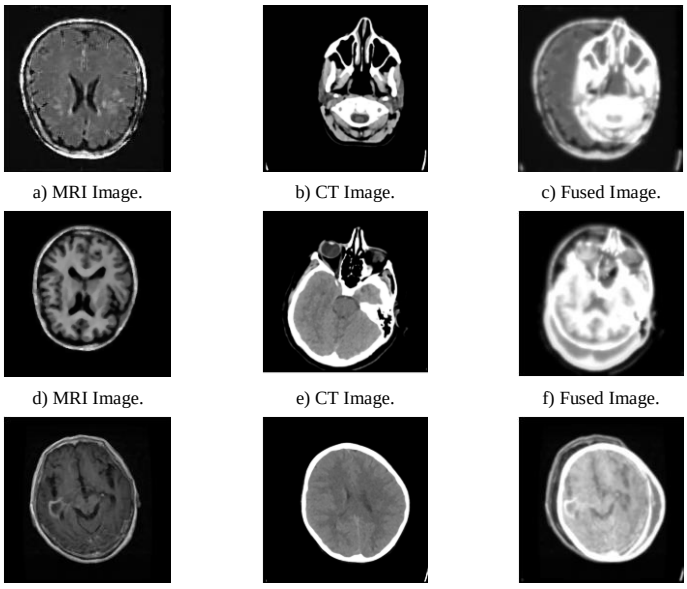
Before performing the fusion process, each input image can be enhanced to highlight the important details and features that are relevant to the fusion process by applying median filtering and histogram equalization. This can enhance the fused image's quality and facilitate interpretation and analysis.

A comparison analysis of the proposed approach with representative conventional and cutting-edge deep learning algorithms for MRI-CT, MRI-PET, and MRI-SPECT image fusion is given in this section. Using the publicly accessible Whole Brain Atlas Harvard dataset, both qualitative and quantitative comparative studies are carried out.

#### 4.3.1. Fusion of MRI and CT Images

For fusing MRI-CT image pairings, the comparison of various techniques is shown in this section. Comparing techniques include Efficient Model Module And Sparse Transformer (EMOST) [26], Multiscale Adaptive Transformer (MATR) [20], efficient fusion network based on dense Res2net and double nonlocal attention models (Res2Fusion) [27], Image Fusion Network Based On Memory Unit (MUFuison) [4], Principal

Component Analysis Network (PCANET) [6], the Parameter Adaptive-Pulse Coupled Neural Network and Non-Subsampled Contourlet Transform (PA-PCNN-NSCT) [8] and Multilevel Bidirectional Feature Interaction Network (MBFINet) [20]. The test set results are used for analyzing the performance of the proposed research. For testing, thirty pairs of MRI-CT images are utilized. The MRI-CT image fusion results are shown in Figure 7. Additionally, the image fusion objective assessment metrics were employed to assess the fused images. According to the existing fusion works, the MUFusion approach can successfully maintain the MRI image's rich texture information. However, it results in edge artifacts. The fused images created using the MATR approach clearly show blurring effects, and the local regions have degraded details. The MRI image's structural information can be effectively characterized by the MBFINet and Res2Fusion approaches, but the CT image's features cannot be adequately presented by them. The fused image is not sufficiently clear since the PCANET approach loses significant texture information. Although the general contour is preserved by the IDPCNN approach, the local magnified region lacks clear texture information. The MRI and CT images' significant features are retained by the PA-PCNN-NSCT technique, however, some detailed information is lost. Furthermore, the DSAGAN adds excessive amounts of extraneous noise, which reduces the fused image's clarity. Using the proposed fusion approach, The MRI and CT images' key features are successfully fused while maintaining the structures and contours. Additionally, the fused image's textures and details produced by the proposed MMIF approach are the clearest.



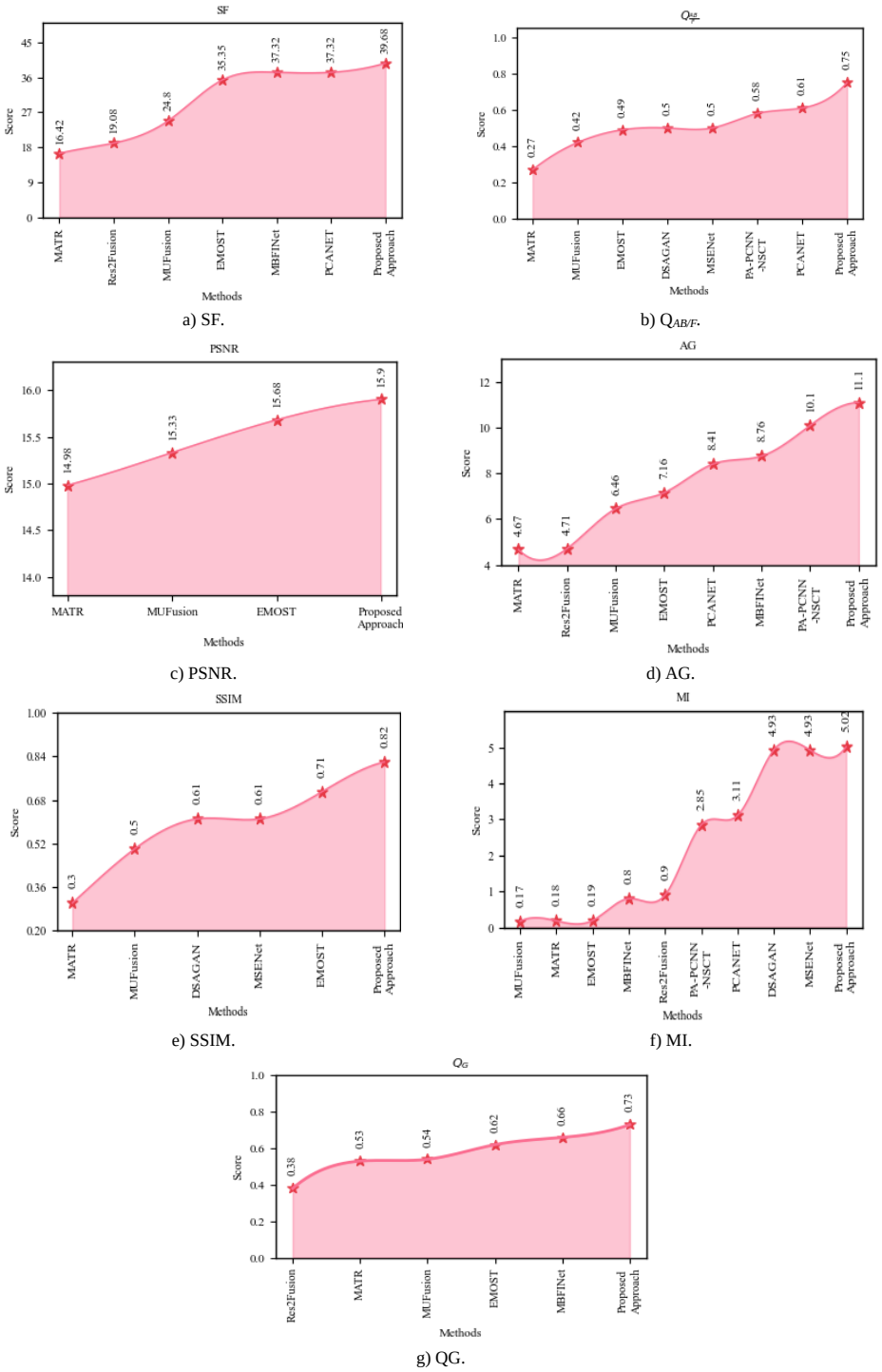
a) MRI Image. b) CT Image. c) Fused Image.
d) MRI Image. e) CT Image. f) Fused Image.
g) MRI Image. h) CT Image. i) Fused Image.

Figure 7. The MRI-CT image pairs' fusion results.

Figure 8 compares the existing fusion methods with the results of MRI-CT image pairings' quantitative evaluation. For each test image, the best fusion results are efficiently produced by the proposed approach in terms of Edge-based Similarity Measure (QAB/F), Spatial Frequency (SF), Peak Signal-to-Noise Ratio (PSNR), Average Gradient (AG), Structural Similarity Index Measure (SSIM), Mutual Information (MI), and Gradient-based Metric (QG). On certain test images, our proposed MMIF approach yields the best results for the AG and MI measures. The typical evaluation outcomes of several techniques on MRI-CT image pairs are shown in Table 5. The proposed approach outperforms the EMOST method across all performance measures. Consequently, the proposed MMIF approach outperforms recent existing methods in terms of qualitative and quantitative fusion performance. The fact that the proposed fusion method achieves high PSNR is indicative of lower noise levels, an increased relationship between the source and different images, and extremely strong information about the association between the fused and original images. The proposed MMIF approach fuses MRI and CT images in an average of 0.221 seconds.

Table 5. The mean values of seven assessment measures of several fusion methods across MRI-CT image pairs.

| Reference | Methods | SF | $Q_G$ | MI | AG | SSIM | $Q_{AB/F}$ | PSNR |
|---|---|---|---|---|---|---|---|---|
| **Wang *et al.* [26]** | EMOST | 35.3573 | 0.6245 | 0.1983 | 7.1684 | 0.7135 | 0.4933 | 15.6856 |
| **Tang *et al.* [24]** | MATR | 16.4298 | 0.5311 | 0.1806 | 4.6715 | 0.3024 | 0.2703 | 14.9843 |
| **Wang *et al.* [27]** | Res2Fusion | 19.0822 | 0.3801 | 0.9054 | 4.7169 | - | - | - |
| **Cheng *et al.* [3]** | MUFusion | 24.8076 | 0.5441 | 0.1792 | 6.468 | 0.5082 | 0.4254 | 15.3391 |
| **Ghandour *et al.* [6]** | PCANET | 37.32781 | - | 3.11297 | 8.419538 | | 0.614455 | - |
| **Ibrahim *et al.* [8]** | PA-PCNN-NSCT | - | - | 2.8512 | 10.103 | - | 0.5819 | - |
| **Shi *et al.* [20]** | MBFINet | 37.3247 | 0.6643 | 0.8078 | 8.7659 | - | - | - |
| **Proposed approach** | MMIF approach | 39.6863 | 0.7362 | 5.0291 | 11.1062 | 0.8264 | 0.7524 | 15.9050 |



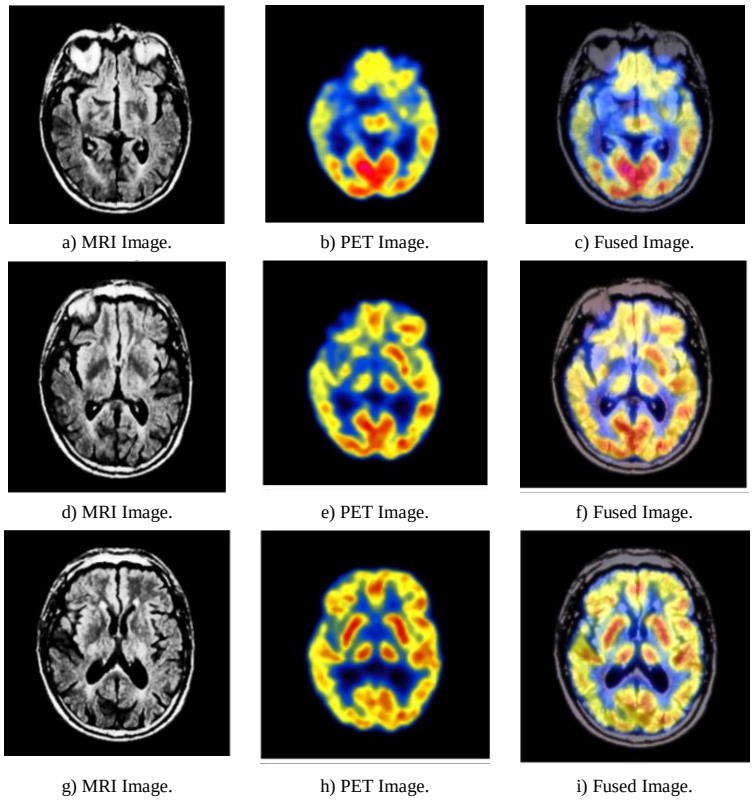

a) SF. b) $Q_{AB/F}$. c) PSNR. d) AG. e) SSIM. f) MI. g) QG.

Figure 8. Performance comparison for MRI and CT image fusion.

### 4.3.2. Fusion of MRI and PET images

For MRI-PET image fusion, the experimental results are shown and discussed in this section. The proposed approach's performance is examined using the test set of MRI-PET image pairs. Figure 9 displays the proposed fusion results for MRI and PET images. The proposed research compared with recent existing fusion works such as EMOST [26], MATR [20], Res2Fusion [27], MUFusion [3], PCANET [6], PA-PCNN-NSCT [8], and MBFINet [20]. In the final fused images, the poor spatial details are produced by the DDCGAN from the comparison analysis. Rich texture information is well preserved by the MUFusion approach. Although the MATR approach can efficiently maintain color information, certain fused images can appear blurry as a result of it. Low brightness in the PCANET fused images significantly impairs the ability to perceive texture details. Although the MBFINet approach does an adequate task of preserving detailed information, the fusion results contain extraneous noise. Spatial distortion is produced at the edge region by the Res2fusion method. The MRI image's detailed information is not properly characterized by the IDPCNN approach, and the blurring effect is visible in

the fused images. The structural information is effectively preserved by the proposed approach without adding noise or distorting color when compared to existing fusion methods.



a) MRI Image. b) PET Image. c) Fused Image.
d) MRI Image. e) PET Image. f) Fused Image.
g) MRI Image. h) PET Image. i) Fused Image.

Figure 9. The MRI-PET image pairs' fusion results.

Fusion results of MRI-PET image pairings' quantitative evaluation with several fusion methods are displayed in Figure 10. The proposed approach works better than the existing fusion methods in terms of $Q_G$, MI, SSIM, AG, PSNR, SF, and $Q_{AB\backslash F}$. The statistical assessment results' average values for MRI-PET image pairs are shown in Table 6. The proposed approach outperforms the most recent EMOST fusion method as per the evaluation values. The proposed approach produced the maximum results for SF, MI, and SSIM. In general, the proposed fusion approach performs better when combining MRI and PET images. Within the pair of image quality indices ($Q_G$ and AG), the proposed algorithm obtained the greatest ranking. The finding that the proposed algorithm's AG index is noticeably greater than the AG indices of the other algorithms is especially important. This result implies that, in comparison to images produced by the other nine compositing algorithms, those produced by the proposed method show enhanced sharpness, contrast, and brightness. The outcomes for the three indices (SSIM, MI, and $Q_{AB/F}$) that were used to assess how much information was preserved in the composite images agreed with the conclusions that had been previously addressed. Concerning these indices, the proposed MMIF approach performed best, but the models EMOST, DSAGAN, PCANET, MBFINet, and IDPCNN had marginally lower scores. However, for these three indications, the models MSENet, MATR, Res2Fusion, and MUFusion showed noticeably lower results. This outcome shows that, in terms of maintaining image details, the proposed model performs better than the other models. The proposed MMIF approach fuses MRI and PET images in an average of 0.051 seconds.

Table 6. The mean values of seven assessment measures of several fusion methods across MRI-PET image pairs.

| Reference | Methods | SF | $Q_G$ | MI | SSIM | AG | $Q_{AB/F}$ | PSNR |
|---|---|---|---|---|---|---|---|---|
| Wang *et al*. [26] | EMOST | 33.7913 | 0.5577 | 0.5051 | 0.7213 | 8.3689 | 0.5845 | 15.2748 |
| Tang *et al*. [24] | MATR | 15.7889 | 0.5272 | 0.496 | 0.3256 | 4.7864 | 0.3437 | 16.1177 |
| Wang *et al*. [27] | Res2Fusion | 21.432 | 0.5502 | 0.9377 | - | 7.1876 | - | - |
| Cheng *et al*. [3] | MUFusion | 23.6499 | 0.4924 | 0.5056 | 0.5041 | 7.0929 | 0.5026 | 15.3141 |
| Ghandour *et al*. [6] | PCANET | 26.02398 | - | 2.32974 | - | 7.51449 | 0.680409 | - |
| Ibrahim *et al*. [8] | PA-PCNN-NSCT | - | - | 3.01 | - | 8.5367 | 0.6805 | - |
| Shi *et al*. [20] | MBFINet | 38.4875 | 0.7602 | 0.9566 | - | 10.9738 | - | - |
| Proposed | MMIF approach | 41.0832 | 0.8513 | 5.9432 | 0.8093 | 11.1883 | 0.7896 | 17.5323 |

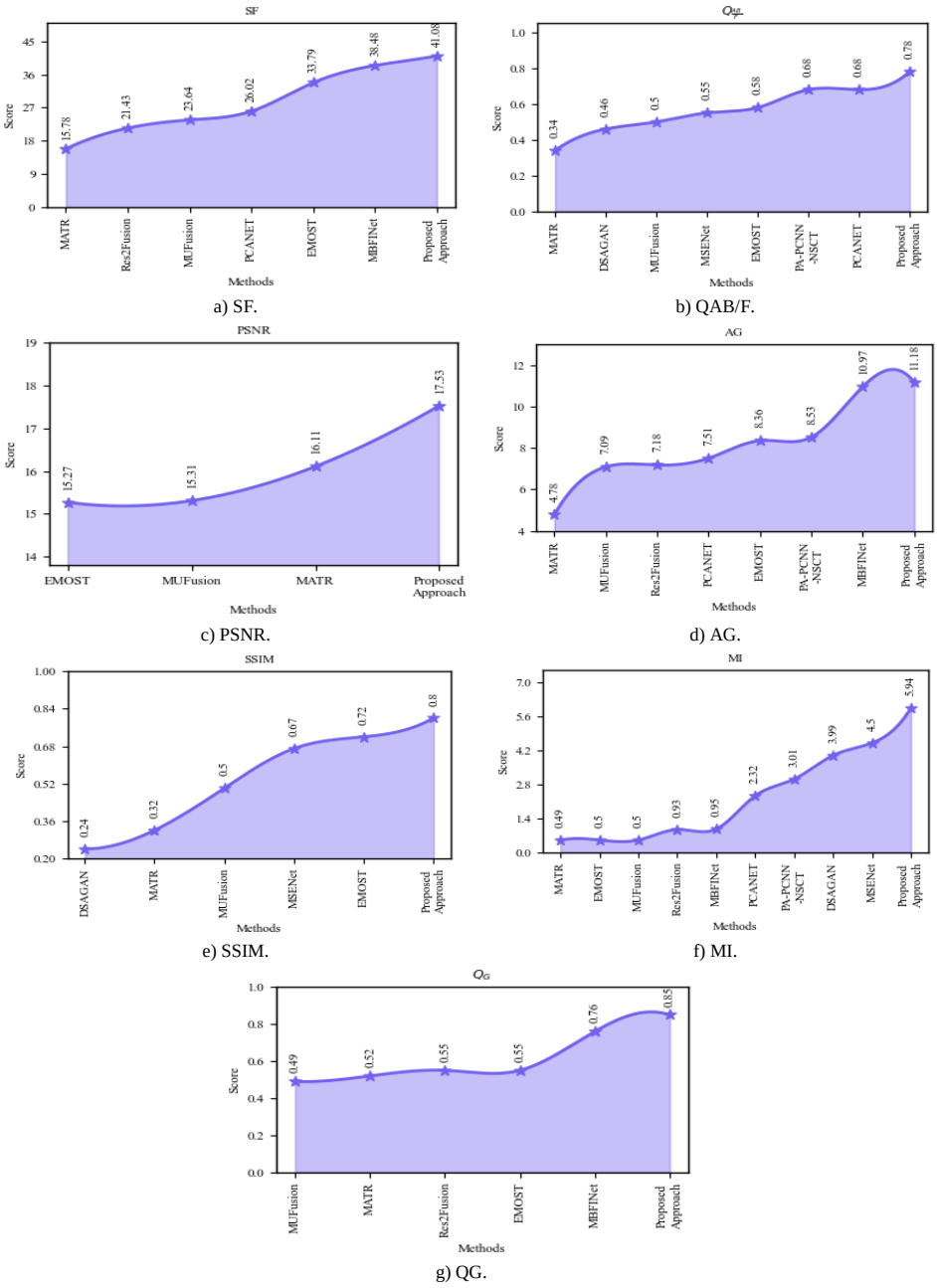
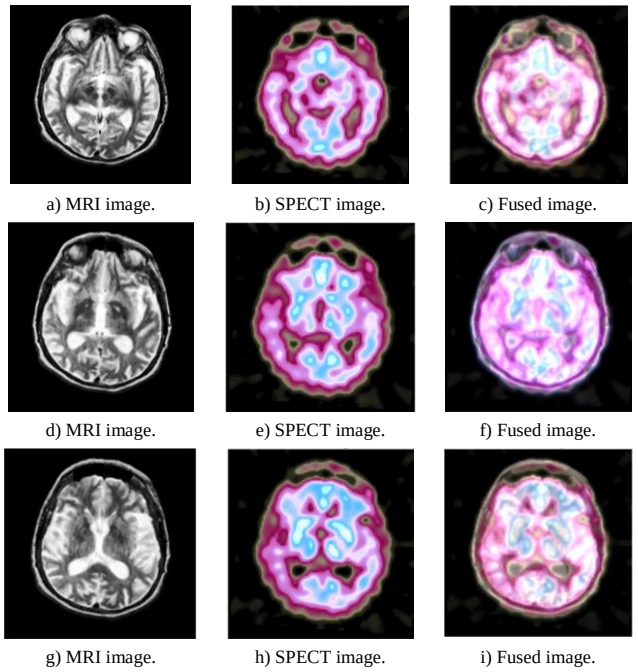



Figure 10. Performance comparison for MRI and PET image fusion.

### 4.3.3. Fusion of MRI and SPECT Images

The experimental results on MRI-SPECT image pairs are shown and discussed in this section. The proposed approach's performance is examined using the test set of MRI-SPECT image pairs. The proposed fusion results for MRI and SPECT are shown in Figure 11. This section compares the proposed approach for fusing SPECT and MRI images with other SOTA methods. The comparison methods include EMOST [26], MATR [20], Res2Fusion [27], MUFusion [3], PCANET [6], PA-PCNN-NSCT [8], and MBFINet [20]. From the existing fusion works, the functional and structural information is not well preserved by the DDCGAN technique. The MUFusion approach may effectively maintain texture details and color information as compared to the DDCGAN method. When applied to edge regions, the MATR technique causes some blurring. Furthermore, the fused images produced by MATR have an excessively bright color. The MSENet approach produces fused images that are excessively blurry. The MRI image contains distorted information as a result of Res2Fusion's inability to maintain the crucial original content. In the dark region, the local information is lost by the PA-PCNN-NSCT method. The MBFINet creates fused images that are less vibrant and struggle to maintain edge details. When compared to the other approaches, ours is better at preserving both structural and functional information, making it more useful for later medical needs. Our approach, however, does not allow the CT image to interfere and preserves more texture information from the MRI image. By contrast, the proposed approach more effectively maintains dense structures in the SPECT source image and edge details in the MRI source image.



a) MRI image. b) SPECT image. c) Fused image.
d) MRI image. e) SPECT image. f) Fused image.
g) MRI image. h) SPECT image. i) Fused image.

Figure 11. The MRI-SPECT image pairs' fusion results.

The proposed approach's performance is examined using the test set results. Thirty pairs of fused images are chosen to assess the proposed research. For MRI-SPECT image fusion, the statistical assessment results of the proposed fusion methods with various existing fusion methods are given in Table 7. Test set MRI-SPECT image pairs of fusion images are chosen to assess the proposed method's efficacy as shown from a visual comparison, as illustrated in Figure 12. As seen in this figure, our proposed approach outperforms the others in terms of $Q_G$, MI, SSIM, AG, PSNR, SF, and $Q_{AB/F}$. When compared to previously published methods, the metric $Q_{AB/F}$ clearly shows the superior performance of our proposed approach, especially concerning information and edge preservation in the generated images. $Q_G$ shows that more texture details are preserved in our fused image. With all factors considered, our proposed fusion approach performs better overall in both quantitative and qualitative evaluations of the MRI-SPECT fusion challenge. The proposed MMIF approach fuses MRI and SPECT images in an average of 0.052 seconds.

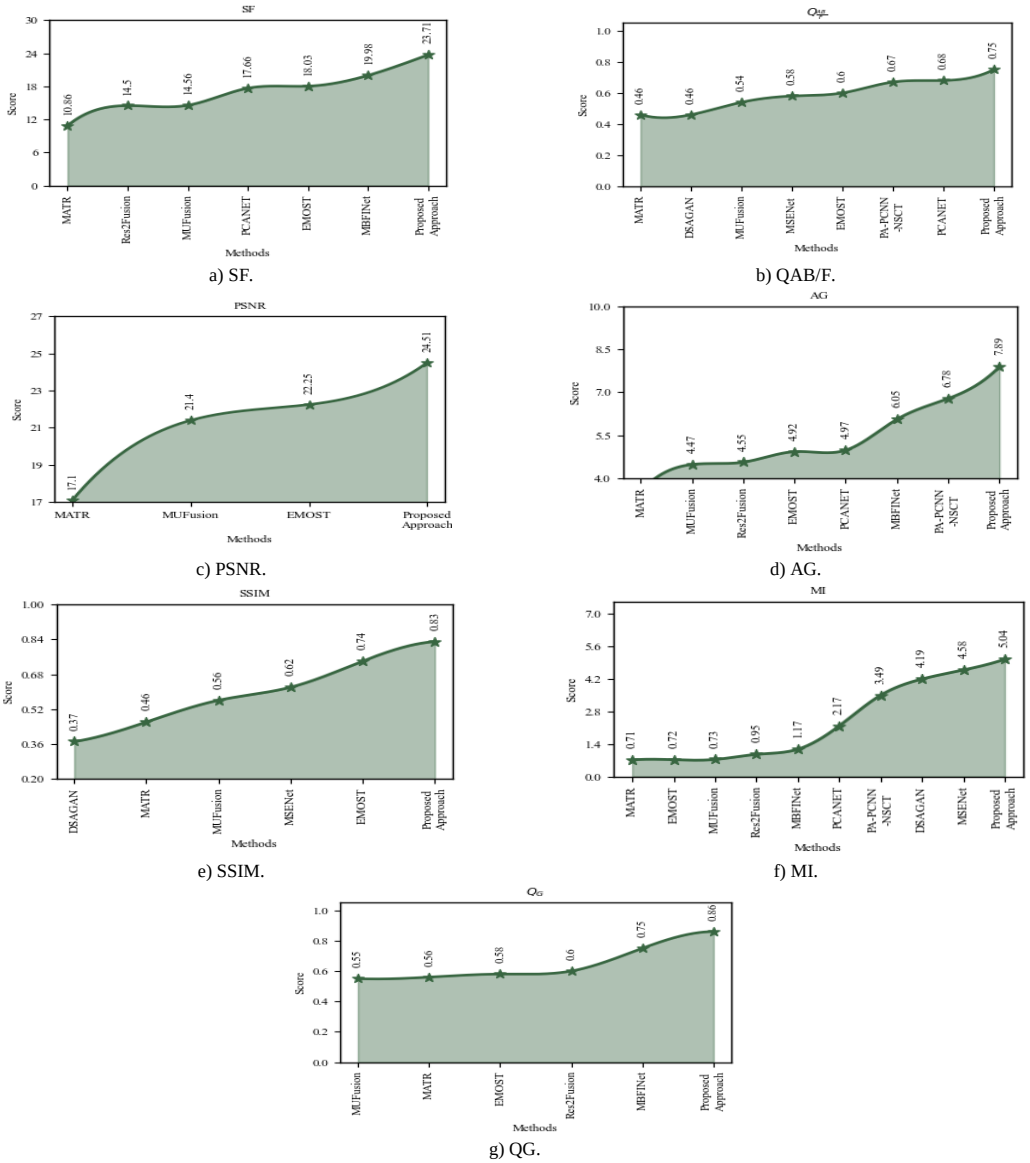



Figure 12. Performance comparison for MRI and SPECT image fusion.

## 4.4. Discussion

In this research, we have used separate effective deep learning models for extracting multi-level features. Low-level features capture fine details, while high-level features capture semantic information. Combining both can enhance the overall quality of the fused image. The proposed MMIF algorithm outperforms the other nine algorithms on almost all average measures, as shown by Tables 5, 6, and 7. The proposed approach nevertheless gives better results even though it only infrequently produces values that are marginally higher than those obtained using other techniques. The proposed MMIF approach outperformed the nine comparative techniques in terms of overall performance. The performance measures $Q_G$, MI, SSIM, AG, SF, PSNR, and $Q_{AB/F}$, measurements are shown in Tables 8, 9, and 10. The fusion outcome performs better when the above-mentioned measures have high values. One purpose of image fusion is to increase comprehensive, adequate, and relevant details, producing a product best suited for human visualization. For fusion approaches to be effective, objective evaluation and visual analysis are equally important. The fusion results of the proposed MMIF approach are shown in Figures 7, 9, and 11 to demonstrate the efficacy of the proposed research. Images show that for all medical dataset pairs, the fusion outputs of our method perform better than alternative fusion methods. The existing fusion methods produced lower values for evaluation measures. To assess the effectiveness and show the resilience of the proposed MMIF approach, we practically assess the impact of transformation results in the QG, MI, SSIM, AG, PSNR, SF, and QAB/F metrics for three medical sets of medical image pairs. We examined the performance of our fusion approach for different evaluation metric values.

Table 7. The mean values of seven assessment measures of several fusion methods across MRI-SPECT image pairs.

| Reference | Methods | SF | $Q_G$ | MI | SSIM | AG | $Q_{AB/F}$ | PSNR |
|---|---|---|---|---|---|---|---|---|
| Wang *et al*. [26] | EMOST | 18.0355 | 0.5878 | 0.7259 | 0.7442 | 4.9279 | 0.6034 | 22.2516 |
| Tang *et al*. [24] | MATR | 10.8671 | 0.5691 | 0.7183 | 0.4658 | 3.526 | 0.4612 | 17.1012 |
| Wang *et al*. [27] | Res2Fusion | 14.5082 | 0.6045 | 0.9526 | - | 4.5531 | - | - |
| Cheng *et al*. [3] | MUFusion | 14.5663 | 0.5595 | 0.7303 | 0.5648 | 4.4729 | 0.545 | 21.4031 |
| Ghandour *et al*. [6] | PCANET | 17.66169 | - | 2.17266 | - | 4.97301 | 0.68963 | - |
| Ibrahim *et al*. [8] | PA-PCNN-NSCT | - | - | 3.4902 | - | 6.7863 | 0.6799 | - |
| Shi *et al*. [20] | MBFINet | 19.9801 | 0.752 | 1.1725 | - | 6.0554 | | - |
| Proposed | MMIF approach | 23.7125 | 0.864 | 5.0431 | 0.8341 | 7.8972 | 0.7593 | 24.5127 |

Table 8. Quantitative results obtained with a combination of different blocks in MRI-CT image pairs.

| Pre-processing (Median filtering+Histogram Equalization) | Low-level feature extraction (IGoogLeNet) | High-level feature extraction (MDenseNet-201) | Feature fusion (Soft Attention) | Feature Reconstruction (Modified Resblock) | SSIM | PSNR | MI | SF | AG |
|---|---|---|---|---|---|---|---|---|---|
| | ✓ | ✓ | ✓ | ✓ | 0.6314 | 13.507 | 4.231 | 30.5491 | 9.2091 |
| ✓ | | ✓ | ✓ | ✓ | 0.4432 | 12.572 | 3.1093 | 27.7654 | 7.7262 |
| ✓ | ✓ | | ✓ | ✓ | 0.3964 | 10.761 | 3.735 | 26.5931 | 6.9282 |
| | | ✓ | ✓ | ✓ | 0.2452 | 9.632 | 2.843 | 24.2626 | 4.2601 |
| | ✓ | | ✓ | ✓ | 0.2875 | 8.016 | 2.7851 | 23.8915 | 4.2006 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.8264 | 15.905 | 5.0291 | 39.6863 | 11.1062 |

Table 9. Quantitative results obtained with a combination of different blocks in MRI-PET image pairs.

| Pre-processing (Median filtering+Histogram Equalization) | Low-level feature extraction (IGoogLeNet) | High-level feature extraction (MDenseNet-201) | Feature fusion (Soft Attention) | Feature Reconstruction (Modified Resblock) | SSIM | PSNR | MI | SF | AG |
|---|---|---|---|---|---|---|---|---|---|
| | ✓ | ✓ | ✓ | ✓ | 0.6415 | 15.618 | 4.132 | 36.6503 | 10.1102 |
| ✓ | | ✓ | ✓ | ✓ | 0.4523 | 13.686 | 3.0183 | 30.8563 | 8.6371 |
| ✓ | ✓ | | ✓ | ✓ | 0.2975 | 11.652 | 3.675 | 27.4045 | 7.047 |
| | | ✓ | ✓ | ✓ | 0.2254 | 10.743 | 2.954 | 26.3637 | 7.3712 |
| | ✓ | | ✓ | ✓ | 0.2966 | 9.107 | 2.895 | 25.9807 | 6.3127 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.8093 | 17.5323 | 5.9432 | 41.0832 | 11.1883 |

Table 10. Quantitative results obtained with a combination of different blocks in MRI-SPECT image pairs.

| Pre-processing (Median filtering+Histogram Equalization) | Low-level feature extraction (IGoogLeNet) | High-level feature extraction (MDenseNet-201) | Feature fusion (Soft Attention) | Feature Reconstruction (Modified Resblock) | SSIM | PSNR | MI | SF | AG |
|---|---|---|---|---|---|---|---|---|---|
| | ✓ | ✓ | ✓ | ✓ | 0.7227 | 23.500 | 4.872 | 21.5792 | 5.0113 |
| ✓ | | ✓ | ✓ | ✓ | 0.6764 | 22.742 | 3.546 | 20.6991 | 4.5490 |
| ✓ | ✓ | | ✓ | ✓ | 0.6096 | 20.884 | 3.886 | 20.6126 | 4.158 |
| | | ✓ | ✓ | ✓ | 0.4478 | 20.968 | 3.0509 | 19.2788 | 3.2901 |
| | ✓ | | ✓ | ✓ | 0.3188 | 19.8018 | 2.654 | 19.8914 | 3.4048 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.8341 | 24.5127 | 5.0431 | 23.7125 | 7.8972 |

Measuring objectively in terms of metrics, the proposed method outperforms earlier comparable methods concerning the information transfer rate for all medical images. Furthermore, in comparison to

alternative methodologies, the average value of assessment metrics for the proposed method is greater for each of the three pairings of medical datasets. Compared to existing fusion procedures, it was observed that the non-reference-based metrics values of the proposed method are highly significant. Comparable methods perform so much less than our proposed MMIF approach, which performs similarly visually. Furthermore, our approach becomes more apparent, showing the least amount of information loss and the highest amount of information transfer. In contrast, in other comparing approaches, fused images show higher levels of noise and information loss.

The novel network model proposed in this research combines the benefits of DenseNet-201, IGoogLeNet, and the Soft Attention mechanism. The model improves the accuracy and efficiency of image processing by combining the benefits of each technique efficiently and drastically lowering the chance of artifacts and edge blurring. The application of the proposed method has proven successful, producing images that are rich in information by combining several modalities of medical imaging, offering more thorough and precise reference data for clinical diagnosis.

### 4.5. Ablation Study

We used test datasets to perform tests on all three sets of image pairs (MRI-CT, MRI-SPECT, and MRI-PET) in order to evaluate the effect of different blocks on the model's effectiveness. For these experimental assessments, three example image pairs were chosen. These images represent various modalities, each displaying unique information relating to the tumor. The main goal was to highlight differences between tumor locations while preserving the fine details from the original maps in the fusion results. Various combinations of the pre-processing, feature extraction, and fusion blocks were used in the adopted ablation studies. Ablation tests on several combinations of proposed modules on MRI-CT, MRI-PET, and MRI-SPECT image pairings are displayed in Tables 8, 9, and 10. According to ablation experiment results for image pairs, all measures perform significantly less well, and the final fusion process takes longer without the pre-processing step. Additionally, the fusion results are marginally impacted by the pre-processing step's absence. Without low-level feature extraction (IGoogleNet), the obtained fused image lacked sharpness and fine details, leading to blurred or less visually informative results. Metrics like SF and AG showed significant reductions, reflecting poorer preservation of local features. Without high-level feature extraction (MDenseNet-201), the fused image lacked meaningful integration of high-level semantics, leading to less interpretable results. We have obtained lower results for SSIM and MI metrics because the fused image needs to retain the meaningful content of the source images. When both low-level and high-level feature extraction are omitted, PSNR is reduced due to noise and loss of fidelity in the reconstructed image, the MI is decreased indicating less information transfer from the source images to the fused image, the final image may appear blurred, less detailed, or semantically incoherent. The proposed research achieved higher performance measures and superior fusion results by utilizing the complete combination of all modules.

### 4.6. Efficiency and Limitations

Based on the experiment analysis, each modal dataset was used to calculate the speed of the proposed approach. For MRI-PET and MRI-SPECT image fusions, the proposed MMIF method requires longer processing times. These findings indicate that, except for the fusion approach based on sparse representations, the majority of deep learning-based fusion techniques are faster than conventional fusion techniques. The proposed approach is quicker than the conventional approaches, but it is a little slower than the other approaches since the fusion of grayscale images happens more slowly than that of color images because 60% of the channels in the grayscale image must be processed, compared to 10% in the color image. Furthermore, we have shown the fusion results on three group datasets, showing that the grayscale fusion effect outperforms the color image fusion both subjectively and quantitatively. It is evident that the important PET/SPECT data are retained, however vibrant colors cover out some of the MRI's finer details. Since just Network processing is done on the Y channel of the YCbCr space in the color image fusion instance of the proposed MMIF approach, extra processing of color information is not required. Thus, color processing and MRI information enhancement will be taken into consideration in future work to better combine color with other texture data.

## 5. Conclusions

The proposed efficient deep learning-based multi-level feature extraction network has been utilized to perform MMIF. First, two distinct and effective deep learning models, such as IGoogLeNet and MDenseNet-201, are used to extract the multi-level features, including low-level and high-level features in the feature extraction phase. Then the Softmax-based soft attention fusion mechanism is used to fuse the high-level and low-level features and features in the feature fusion phase. Finally, the fused high-level and low-level features are combined using the Modified Resblock module to form the final fused image in the image reconstruction phase. In comparison to cutting-edge methods, the proposed MMIF fusion approach with its SA fusion mechanism performed better on seven performance metrics, including $Q_G$, MI, SSIM, AG, PSNR, SF, and $Q_{AB/F}$. The

results of the proposed MMIF approach with SA fusion mechanism and effective multi-level feature extraction networks showed minimized noise, maximum mutual information and high structural similarity in the fused image. Additionally, the best results in terms of $Q_G$, AG, and SF values were obtained using the proposed MMIF approach. According to the subjective evaluation, the proposed MMIF approach with SA fusion resulted in more visually understandable images for medical professionals. In terms of certain performance metrics, no other task has produced better results. The effectiveness of the proposed fusion network was verified by comparing it with nine state-of-the-art approaches using three modal datasets. Our proposed MMIF fusion approach exhibited enhanced performance compared with all comparator models such as EMOST, DSAGAN, MATR, MSENet, Res2Fusion, MUFusion, PCANET, PA-PCNN-NSCT, and MBFINet.

Future studies may investigate utilizing the whole fusion block's MATNet block since the promising results are produced by the attention-based networks in highlighting relevant features. To improve the association between an object's appearance and motion, the two-stream encoder for motion-attentive representations is created using MATNet by integrating the MAT block. Better fusion results are achieved by extracting selective representations from multi-level encoder data through the use of a bridge network.

## Declaration

### Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

**Availability of data and material**: data will be available when requested.

**Funding**: no funding was received to assist with the preparation of this manuscript.

**Authors' contributions**: the author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

**Ethics approval**: this material is the authors' own original work, which has not been previously published elsewhere. The paper reflects the authors' own research and analysis in a truthful and complete manner.

## References

[1] Ahamed B., Baskar R., and Nalinipriya G., "Enhanced Brain Tumor MRI Scan Reconstruction via the EI-Fusion-Net Model," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 4, pp. 1-23, 2024. DOI: 10.22266/ijies2024.0831.53

[2] Avci D., Sert E., Ozyurt F., and Avci E., "MFIF-DWT-CNN: Multi-focus Image Fusion Based On Discrete Wavelet Transform with Deep Convolutional Neural Network," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 10951-10968, 2024. https://doi.org/10.1007/s11042-023-16074-6

[3] Cheng C., Xu T., and Wu X., "MUFusion: A General Unsupervised Image Fusion Network Based on Memory Unit," *Information Fusion*, vol. 92, no. 1, pp. 80-92, 2023. https://doi.org/10.1016/j.inffus.2022.11.010

[4] Cheng Y., Fang X., Tang Z., Yu Z., Sun L., and Zhu L., "SDR2Tr-GAN: A Novel Medical Image Fusion Pipeline Based on GAN with SDR2 Module and Transformer Optimization Strategy," *International Journal of Imaging Systems and Technology*, vol. 34, no. 6, pp. 23208, 2024. https://doi.org/10.1002/ima.23208

[5] Ding W., Geng S., Wang H., Huang J., and Zhou T., "FDiff-Fusion: Denoising Diffusion Fusion Network Based on Fuzzy Learning for 3D Medical Image Segmentation," *Information Fusion*, vol. 1, no. 1, pp. 102540, 2024. https://doi.org/10.1016/j.inffus.2024.102540

[6] Ghandour C., El-Shafai W., El-Rabaie E., and Elshazly E., "Applying Medical Image Fusion Based on a Simple Deep Learning Principal Component Analysis Network," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5971-6003, 2024. https://doi.org/10.1007/s11042-023-15856-2

[7] He D., Li W., Wang G., Huang Y., and Liu, S., "LRFNet: A Real-Time Medical Image Fusion Method Guided by Detail Information," *Computers in Biology and Medicine*, vol. 173, no. 1, pp. 108381, 2024. https://doi.org/10.1016/j.compbiomed.2024.108381

[8] Ibrahim S., El-Tawel G., and Makhlouf M., "Brain Image Fusion Using the Parameter Adaptive-Pulse Coupled Neural Network (PA-PCNN) and Non-Subsampled Contourlet Transform (NSCT)," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 27379-27409, 2024. https://doi.org/10.1007/s11042-023-16515-2

[9] Kahol A. and Bhatnagar G., "Deep Learning-

Based Multimodal Medical Image Fusion," *Data Fusion Techniques and Applications for Smart Healthcare*, vol. 1, no. 1, pp. 251-279, 2024. https://doi.org/10.1016/B978-0-44-313233-9.00017-5

[10] Khan S., Alharbi M., Shah S., and ELAffendi M., "Medical Image Fusion for Multiple Diseases Features Enhancement," *International Journal of Imaging Systems and Technology*, vol. 34, no. 6, pp. 23197, 2024. https://doi.org/10.1002/ima.23197

[11] Kumari B., Nandal A., and Dhaka A., "Breast Tumor Detection Using Multi-Feature Block Based Neural Network by Fusion of CT and MRI Images," *Computational Intelligence*, vol. 40, no. 3, pp. 12652, 2024. https://doi.org/10.1111/coin.12652

[12] Li B., Wang J., Wang B., Shao Z., Li W., Huang J., and Li P., "BMCS-Net: A Bi-Directional Multi-Scale Cascaded Segmentation Network Based on Transformer-Guided Feature Aggregation for Medical Images," *Computers in Biology and Medicine*, vol. 180, pp. 108939, 2024. https://doi.org/10.1016/j.compbiomed.2024.108939

[13] Liu Y., Yu C., Cheng J., Wang Z., and Chen X., "MM-Net: A Mixformer-Based Multi-Scale Network for Anatomical and Functional Image Fusion," *IEEE Transactions on Image Processing*, vol. 33, pp. 2197-2212, 2024. DOI: 10.1109/TIP.2024.3374072

[14] Mishra N. and Dhabal S., "An Improved Hybrid Fusion of Noisy Medical Images Using Differential Evolution-Based Artificial Rabbits Optimization Algorithm," *Multidimensional Systems and Signal Processing*, vol. 35, pp. 1-55. 2024. https://doi.org/10.1007/s11045-024-00889-z

[15] Nehru V. and Prabhu V., "Automated Multimodal Brain Tumor Segmentation and Localization in MRI Images Using Hybrid Res2-UNeXt," *Journal of Electrical Engineering and Technology*, pp. 1-13, 2024. https://doi.org/10.1007/s42835-023-01779-3

[16] Raj S. and Singh B., "SpFusionNet: Deep Learning-Driven Brain Image Fusion with Spatial Frequency Analysis," *Multimedia Tools and Applications*, vol. 83, pp. 82983-83004, 2024. https://doi.org/10.1007/s11042-024-18682-2.

[17] Ramaraj V., Swamy M., and Sankar M., "Medical Image Fusion for Brain Tumor Diagnosis Using Effective Discrete Wavelet Transform Methods," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 70-80, 2024. https://doi.org/10.20473/jisebi.10.1.70-80

[18] Ravi J. and Narmadha R., "Multimodality Medical Image Fusion Analysis with Multi-Plane Features of PET and MRI Images Using ONSCT." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 11, no. 7, pp. 2255684, 2024. https://doi.org/10.1080/21681163.2023.2255684

[19] Roy M. and Mukhopadhyay S., "A DCT-Based Multiscale Framework for 2D Greyscale Image Fusion Using Morphological Differential Features," *The Visual Computer*, vol. 40, no. 5, pp. 3569-3590, 2024. https://doi.org/10.1007/s00371-023-03052-0

[20] Shi K., Liu A., Zhang J., Liu Y., and Chen X., "Medical Image Fusion Based on Multi-Level Bidirectional Feature Interaction Network," *IEEE Sensors Journal*, vol. 24, no. 12, pp. 19428-19441, 2024. DOI: 10.1109/JSEN.2024.3393619

[21] Sinha A., Agarwal R., Kumar V., Garg N., Pundir D., Singh H., Rani R., and Panigrahy C., "Multi-Modal Medical Image Fusion Using Improved Dual-Channel PCNN," *Medical and Biological Engineering and Computing*, vol. 62, pp. 2629-2651, 2024. https://doi.org/10.1007/s11517-024-03089-w

[22] Song W., Zeng X., Abdelmoniem A., Zhang H., and Gao M., "Cross-Modality Interaction Network for Medical Image Fusion," *IEEE Transactions on Consumer Electronics*, vol. 99, pp. 1-1, 2024. DOI: 10.1109/TCE.2024.3412879

[23] Sun Y., Li W., Xing J., Zhang B., Pu D., Liu Q., and Sun Y., "Enhancing Session-Based Recommendations by Fusing Candidate Items," *The International Arab Journal of Information Technology*, vol. 21, no. 6, pp. 1029-1042, 2024. https://doi.org/10.34028/iajit/21/6/7

[24] Tang W., He F., Liu Y., and Duan Y., "MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134-5149, 2022. DOI: 10.1109/TIP.2022.3193288

[25] Tirupal T., Pandurangaia Y., Roy A., Kishore V., and Nayyar A., "On the use of UDWT and Fuzzy Sets for Medical Image Fusion," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 39647-39675, 2024. https://doi.org/10.1007/s11042-023-16892-8

[26] Wang W., He J., and Liu H., "EMOST: A Dual-Branch Hybrid Network for Medical Image Fusion Via Efficient Model Module and Sparse Transformer," *Computers in Biology and Medicine*, vol. 179, pp. 108771, 2024. https://doi.org/10.1016/j.compbiomed.2024.108771

[27] Wang Z., Wu Y., Wang J., Xu J., and Shao W., "Res2Fusion: Infrared and Visible Image Fusion Based on Dense Res2net and Double Nonlocal Attention Models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-12. 2022. DOI: 10.1109/TIM.2021.3139654

[28] Wei Y. and Ji L., "Multi-Modal Bilinear Fusion

with Hybrid Attention Mechanism for Multi-Label Skin Lesion Classification," *Multimedia Tools and Applications*, vol. 83, pp. 65221-65247, 2024. https://doi.org/10.1007/s11042-023-18027-5

[29] Xie X., Zhang X., Tang X., Zhao J., Xiong D., Ouyang L., Yang B., Zhou H., Ling B., and Teo K., "MACTFusion: Lightweight Cross Transformer for Adaptive Multimodal Medical Image Fusion," *IEEE Journal of Biomedical and Health Informatics*, 2024. DOI: 10.1109/JBHI.2024.3391620

[30] Xie Y., Yu L., and Ding C., "CFIFusion: Dual-Branch Complementary Feature Injection Network for Medical Image Fusion," *International Journal of Imaging Systems and Technology*, vol. 34, no. 4, pp. 23144, 2024. https://doi.org/10.1002/ima.23144

[31] Zhang W., Yu L., Wang H., and Pedrycz W., "End-to-End Dynamic Residual Focal Transformer Network for Multimodal Medical Image Fusion," *Neural Computing and Applications*, vol. 36, pp. 11579-11601, 2024. https://doi.org/10.1007/s00521-024-09729-4

[32] Zhao H., Cai H., and Liu M., "Transformer Based Multi-Modal MRI Fusion for Prediction of Post-Menstrual Age and Neonatal Brain Development Analysis," *Medical Image Analysis*, vol. 94, pp. 103140, 2024. https://doi.org/10.1016/j.media.2024.103140

[33] Zhou Y., He K., Xu D., Tao D., Lin X., and Li C., "ASFusion: Adaptive Visual Enhancement and Structural Patch Decomposition for Infrared and Visible Image Fusion," *Engineering Applications of Artificial Intelligence*, vol. 132, pp. 107905, 2024. https://doi.org/10.1016/j.engappai.2024.107905

[34] Zhou Y., Yang X., Liu S., and Yin J., "Multimodal Medical Image Fusion Network Based on Target Information Enhancement," *IEEE Access*, vol. 12, pp. 70851-70869, 2024. DOI: 10.1109/ACCESS.2024.3402965

**Syed Munawwar** is an Assistant Professor of ECE department at Santhiram Engineering College, Nandyal, Affiliated to Jawaharlal Nehru, Technological University Anantapur, Ananthapuramu-515002, Andhra Pradesh, and Research scholar in Jawaharlal Nehru Technological University Anantapur, Ananthapuramu-515002 Andhra Pradesh, India. His research interests include Machine Learning and Signal Processing, Neural Engineering and Biomedical Systems and Robotics and AI.

**Panyam Vuppu Gopi Krishna Rao** is currently working as a Professor in the Department of Electronics and Instrumentation Engineering in Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Affiliated to Jawaharlal Nehru Technological University Anantapur, Ananthapuramu-515002, Andhra Pradesh, India. He received, B. Tech (IT) Mysore University, M.E (Electronics and Control) Satyabama University and Ph.D. (Robust Process Control) JNTUK, Kakinada. His current research interests include Network Analysis, Electric Circuits, Design Thinking for Innovation, and Control Systems.