





A Review of Deep Learning-based Multi-modal Medical Image Fusion



Shailesh Bhosekar¹ , Prabhishek Singh² , Deepak Garg¹ , Vinayakumar Ravi^{5,*} and Manoj Diwakar^{3,4} 

¹School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India, 506371

²School of Computer Science Engineering and Technology, Bennett University, Greater Noida 201310, India

³Department of CSE, Graphic Era Deemed to be University, Dehradun, Uttarakhand, 248002, India

⁴Graphic Era Hill University, Dehradun, Uttarakhand, 248002, India

⁵Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia

Abstract:

Introduction: Medical image fusion combines the data obtained from different imaging modalities such as Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) into a single, informative image that aids clinicians in diagnosis and treatment planning. No single imaging modality can provide complete information on its own. This has led to the emergence of a research field focused on integrating data from multiple modalities to maximize information in a single, unified representation.

Methods: CNN (Convolutional Neural Network) was applied to achieve robust and effective multi-modal image fusion. By delving into the principles and practical applications of this deep learning approach, the paper also provides a comparative analysis of CNN-based results with other conventional fusion techniques.

Results: CNN-based image fusion delivers far better results in terms of qualitative and quantitative analysis when compared with other conventional fusion methods. The paper also discusses future perspectives, emphasizing advancements in deep learning that could drive the evolution of CNN-based fusion and enhance its effectiveness in medical imaging.

Discussion: CNN-based multi-modal medical image fusion proves strong advantages over traditional methods in terms of feature preservation and adaptability. However, challenges such as data dependency, computational complexity, and generalization across modalities persist. Emerging trends like attention mechanisms and transformer models show promise in addressing these gaps. Future work should focus on improving interpretability and clinical applicability, ensuring that deep learning fusion methods can be reliably integrated into real-world diagnostic systems.

Conclusion: Ultimately, this work underscores the potential of CNN-based fusion to improve patient outcomes and shape the future of medical imaging by advancing the understanding of multi-modal fusion.

Keywords: Image fusion, Multi-modal medical image fusion, PET image, CT image, MRI image, Method noise, CNN.

© 2025 The Author(s). Published by Bentham Open.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: <https://creativecommons.org/licenses/by/4.0/legalcode>. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*Address correspondence to this author at the Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia; E-mail: vinayakumarr77@gmail.com

Cite as: Bhosekar S, Singh P, Garg D, Ravi V, Diwakar M. A Review of Deep Learning-based Multi-modal Medical Image Fusion. Open Bioinform J, 2025; 18: e18750362370697. <http://dx.doi.org/10.2174/0118750362370697250630063814>



Received: November 12, 2024

Revised: January 26, 2025

Accepted: February 06, 2025

Published: July 04, 2025



Send Orders for Reprints to
reprints@benthamscience.net

1. INTRODUCTION

In the last few years, multi-modal image fusion has become a much-favored approach in numerous application areas, including medical imaging, remote sensing, surveillance, and autonomous systems. Combining complementary information, fusion of images from several modalities gives a larger perspective of complicated situations. Here, as an example, in medical imaging, CT scans give structural details, MRI gives soft tissue contrast, and together offer images with better diagnostic value. Likewise, in remote sensing, multi-spectral images sample a group of different environmental attributes, a panchromatic image gives high resolution in spatial (spatial) details, and their fusion produces an image that retains both spectral and spatial information. There is no single modality that can give the complete diagnostic details. Therefore, fusion of medical images is preferred. Multi-modal image fusion is the concept of combining data from different imaging sources into a single informative output image to support better analysis and interpretation, and decision making [1]. Driven by improvements in artificial intelligence (specifically deep learning), the landscape of multi-modal image fusion has changed dramatically. Fusion methods in traditional frameworks are based on mathematical models and manual feature extraction methods such as wavelet transform, Principal Component Analysis (PCA) or other statistical approaches. However, the high complexity and variability of data from different modalities present significant limitations to the application of these methods. In contrast, deep learning-based fusion techniques can automatically learn elaborate features and patterns and are well-suited for multi-modal fusion tasks. Particularly, CNNs, Generative Adversarial Networks (GANs), and transformers lend themselves to this task, as they are capable of extracting complex features and relationships across different imaging modalities [2].

Deep learning provides one of the significant advantages for multi-modal image fusion. It can learn complex hierarchical representations automatically. Typically, traditional fusion methods rely on manual intervention to extract and select features, which is a time-consuming and error-prone

effort. Whereas the process is, however, automated in deep learning in a data-driven way, both low-level and high-level features are learned. Although this capability could be useful in many domains, it is particularly important in domains such as medical imaging, where small differences between different modalities can mean big differences in terms of information. As an example, a well-trained deep learning model can bridge individual details of a modality in the fused output directly and maintain them, therefore, as it generates images that are richer in information and appropriate for diagnostic purposes [3]. With CNNs, deep learning models are flexible and can be trained for pixel-level, feature-level and decision-level fusion. On the pixel level, fusion is of raw pixel information of each modality, whereas, in the feature level, fusion is of higher-level features extracted from convolutional layers. In contrast, decision-level fusion reflects the combination of decisions or interpretations of each modality. Because of this flexibility, deep learning models can be used for a variety of fusion tasks at both low levels as imaging and at high levels as interpretation [4]. The last advantage is that the deep learning-based fusion methods are scalable. Deep learning models can perform multispectral image processing quickly upon being trained, and this is valuable for real-time (or near real-time) processing in applications such as autonomous driving and surveillance. In addition, deep learning models can also frequently be fine-tuned for one use case and work well with other types of data or with new imaging conditions. Adaptability is extremely useful in fields such as remote sensing, where environmental conditions change rapidly, leading to a variety of input image quality and characteristics [5]. The example of multi-modal medical image fusion is shown in Fig. (1).

While deep learning based multi modal image fusion has many advantages, it also has a few distinct disadvantages. The big challenge is that a huge amount of labeled data is required to train. The effectiveness of the deep learning model based on CNN and GAN, *etc*, highly depends on a sufficient amount of data for training. In the medical imaging domain, for instance, collecting a multi-modal labeled dataset may be difficult for privacy reasons, limited access to high-quality data and labeling cost [6].

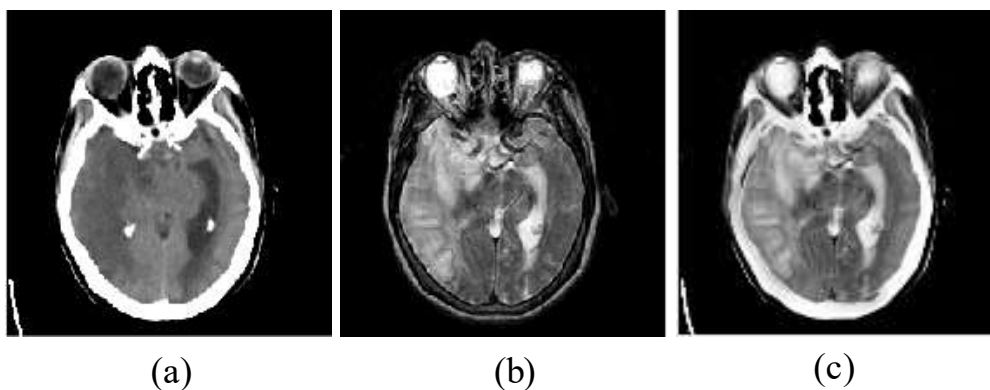


Fig. (1). Example of multi-modal medical image fusion (a) Source image 1— CT image, (b) Source image 2— MRI image, and (c) Fused image.

The second point to be mentioned is that deep learning models are computationally intensive. Training a deep neural network, particularly one with many layers, is quite computationally expensive (on the order of high-performance GPUs or cloud-based computing). However, this computational demand limits easy accessibility of deep learning-based fusion for institutions or researchers with limited resources. Additionally, most deep learning models are treated as 'black boxes', meaning that the internal reasoning behind their decisions cannot be easily interpreted. This lack of transparency is a problem in fields such as medicine, where knowledge of the rationale for a fused image is often needed for clinical decision making. However, interpretability is an active limitation in deep learning-based methods, and researchers are hard at work on interpretability techniques right now [7]. A central challenge in this field is how to design fusion architectures capable of dealing with the heterogeneity of different modalities. The images are of different modalities with large differences in spatial resolution, intensity distribution and structural properties. For example, CT and MRI scans will show you different types of tissues and will have differing types of resolution. It is challenging to develop a deep learning model that can usefully integrate these diverse types of information into a single, coherent output. However, when misalignment and different resolutions are present in inputs, they can produce artifacts in the fused output if not handled [8].

As a result, developing and benchmarking the deep learning-based fusion models can be tricky and requires task-specific metrics [9]. As another significant challenge, model generalization still needs to be resolved. However, the robustness of these models is limited by their sensitivity to generalizing from specific datasets or imaging conditions [10]. For example, a model learned from Medical Images taken from a particular type of MRI machine may not work on Images acquired from a different machine or protocol. However, the lack of generalizability in this regard restricts the use of fusion models for real-world tasks in which variability of imaging conditions is common [11]. There are some persistent problems in the current deep learning-based image fusion. One such issue is that when some fusion techniques tend to produce overly smooth or blurred fused images, fine details get lost. For instance, when models seek to introduce spatial consistency at the expense of modality-specific information, the resulting fused image is poorly sampled, lacking sharpness and possibly containing clinically relevant information [12]. There is a problem with the limited interpretability of fused images. Deep learning models can create high-quality fused images, but the lack of interpretability of deep learning models makes it hard for practitioners to trust or adopt this approach. Medical practitioners may be reluctant to use fused images for diagnosis if they do not know how a fusion process has preserved or changed certain anatomical structures. There are a lot of investigations being conducted to make deep learning models more interpretable, but interpretability still stands as a roadblock for deeper adoption [13]. Furthermore, multimodal datasets require models that can accommodate temporal and spatial inconsistencies. In a dynamic scenario such as autonomous driving, images from different

sensors (*e.g.*, radar, LiDAR, camera) are taken at a slightly different time or from different viewpoints. However, due to this temporal and spatial discrepancy, it will cause challenges in fusion, the fused image can have artifacts or inconsistent areas that may harm the downstream tasks. Such problems are difficult for traditional fusion methods, and although deep learning can provide some answers, more research is needed to develop robust models that can deal with this in real time [14]. Another area of ongoing research is the further development of fusion models suitable for use with the varying quality of data. In practice, input images differ in quality from one image to another, *e.g.*, low noise and good lighting conditions, low noise and low light conditions, *etc.* Automatic selection of fusion algorithms, however, is a task that has not been fully addressed before. Designing fusion algorithms that flexibly adjust their processing depending on the image quality of the input images is a challenging objective [15]. In conclusion, though the multi-modal image fusion is greatly promoted by deep learning, there are still many problems. They include data requirements, computational demands, model interpretability, issues concerning generalizability and robustness. To solve these struggles, further research and innovative work are needed, because deep learning presents one avenue to augment multi-modal image fusion and make it possible in real-world scenarios and various applications [16].

After the introduction, Section 2 briefly describes the theoretical foundations and importance of CNN-based multi-modal medical image fusion. Section 3 describes the results of CNN-based image fusion. Section 4 describes the step-by-step procedure of CNN-based image fusion, and this paper is concluded in Section 5.

2. MATERIALS AND METHODS

Medical imaging has soared in the wake of the introduction of a variety of imaging modalities like CT, MRI, PET, and ultrasound images. All modalities give unique insight into the body, collecting different information, anatomical or functional. If only one imaging modality is available, it is powerful, but it is usually limited to the type of information that it can capture. For example, MRI has great soft tissue contrast but poor bone detail, while CT scans give very good bone detail but are not contrasted in soft tissues. However, to overcome these limitations, multi-modal medical image fusion has been proposed to merge complementary information from multiple imaging modalities to generate a single enriched image. Over recent years, CNNs have become central in multi-modal medical image fusion, providing effective and sophisticated techniques to fuse images to maintain important details and improve the diagnostic accuracy [17].

2.1. Role of CNN-Based Multi-Modal Medical Image Fusion

CNN-based multi-modal medical image fusion plays an important role in combining multiple modalities to offer clinicians a richer insight into a patient's anatomy and physiology. In comparison to more traditional fusion methods that rely on manually extracting features and fixed rules, CNNs learn the fusion task from data, deriving the relevant

features, as well as the fusion, in a data-driven way. The ability to learn and adapt gained from CNNs makes them particularly well suited for medical applications where subtle differences in the images have a critical impact on the diagnosis [18].

CNN-based fusion is also used in other medical-related fields. For example, in oncology, where metabolic information from PET is fused with anatomical detail from MRI, to accurately localize tumors. CNNs can automatically fuse these modalities to provide a detailed image that can assist oncologists in their treatment planning and monitoring [19]. When using neuroimaging technologies and combining MRI and functional MRI (fMRI) via CNNs, neurologists can request an assessment of brain activity and structure from the same image, thus improving diagnostics of diseases such as epilepsy and Alzheimer's disease. In addition, fusion of CT and MRI images in cardiovascular imaging enables examination of the structure and function of the heart, supporting the comprehensive cardiac evaluation [20].

2.2. Significance of CNN-Based Multi-Modal Medical Image Fusion

Since the fusion of high-resolution CNN-based images can offer accurate and high-resolution images employing both the anatomical and functional information with regard to diagnostics, it has become an important issue in the field of medical image fusion. Unlike the traditional flat layers which are used in the feedforward neural network, where it can learn from only predefined features, or merely surface textures, CNNs with deep layers and convolutional filters can learn from complex patterns and textures that exist in medical images and help with distinguishing important details from noise. Importantly, this capability allows to produce fused images that both visually and diagnostically inform with respect to medical conditions for improved interpretation by the physician [21]. One particularly interesting advantage of CNNs is the capability of doing pixel-level and feature-level fusion. In pixel-level fusion, information from each pixel in the source images is fused, preserving fine details, while in feature-level fusion, higher-level features are aggregated, which captures more abstract and diagnostically important features. The dual capability of CNN-based fusion models ensures maintaining important diagnostic information such as tumor boundaries, tissue abnormalities, and vascular structures. Thereby keeping the fused image both precise and informative [22].

Beyond image quality, CNN-based fusion has significance to medical practitioners' workflow. Using CNN, the fusion process is automated, which makes the interpretation and combination of multiple modalities less time-consuming compared with the manual process. The most significant benefit of this is in time savings, which are especially crucial during emergency cases, where diagnosis must happen as soon as possible. Furthermore, the decrease in manual intervention also decreases variability and subjectivity in diagnostics, resulting in more uniform and dependable diagnostic results across various healthcare environments [23].

2.3. Impact of CNN-Based Multi-Modal Medical Image Fusion on Healthcare

CNN-based multi-modal image fusion has a high impact on healthcare, especially in diagnosing rapidly and accurately and personalizing the treatment. CNN-based fusion can create images that contain far greater anatomical and functional information than current techniques, allowing earlier and more accurate diagnoses with better patient outcomes. For example, in the field of cancer diagnosis, the combined fused images of PET and MRI can lead to early detection of tumors and, consequently, to achieving timely intervention. In addition, like neuroimaging, fusing MRI and fMRI gives a comprehensive view of the brain, helpful in precisely localizing functional defects or structural abnormalities, in conditions that require surgical management [24]. Another major impact of CNN-based fusion is on the contribution of precision medicine. In the process of personalized treatment planning, it is extremely important to consider the personal anatomical and physiological properties of the patient. The detailed and personalized view seen in the fused images allows clinicians to tailor treatments to the patient more specifically. In radiation therapy, for example, fused images are particularly useful because they allow precision striking of tumors with radiation while sparing surrounding healthy tissue. As a result, CNN-based fusion helps make treatments less risky and more effective, lowering the risk of complications and improving patients' quality of life [25].

Furthermore, CNN-based fusion models have proved to be a must in developing the medical imaging field. The development and refinement of existing CNN-based fusion techniques help provide better imaging protocols, making it possible to study disease, track disease progression and gauge treatment response. On the other hand, this advancement results in the perpetual evolution of diagnosis tools, and CNN-based fusion methods shape the new standards in medical imaging. The implementation of these cutting-edge fusion techniques may also dictate future healthcare policy, including diagnostic procedure guidelines and treatment planning [26].

2.4. Challenges and Limitations in CNN-Based Multi-Modal Medical Image Fusion

2.4.1. Difficulty in Annotated Dataset Acquisition

- Large, annotated datasets are required to train CNN models effectively.
- Privacy issues, regulatory constraints, and a lack of multi-modal datasets hinder data acquisition.
- The expertise and time needed to annotate images make it challenging to create high-quality datasets [27].

2.4.2. Generalization Issues

- Without proper data, CNN models may fail to generalize, leading to inconsistent and inaccurate fusion results [27].

2.4.3. Computational Complexity

- High computational power and memory are required for training CNN models on large medical datasets.
- Resource demands (*e.g.*, GPUs or cloud-based solutions) may not be accessible to all healthcare institutions.
- Real-time applications are limited due to the high computational demand, restricting integration into clinical workflows in resource-limited settings [28].

2.4.4. Lack of Interpretability

- CNN architectures are complex and often viewed as “black boxes.”
- Clinicians may struggle to understand the rationale behind fusion decisions.
- Transparency and trust in fused images are critical for accurate diagnosis, which is hindered by the lack of interpretability [29].

2.4.5. Non-Generalizability Across Protocols

- CNN models trained on images from specific scanners or protocols may not work well across different equipment or imaging protocols.
- Variability in medical imaging protocols can create challenges in multi-center studies or deployments across diverse healthcare facilities [30, 31].

2.4.6. Skepticism from Healthcare Providers

- The lack of interpretability may lead to skepticism among healthcare providers, reducing motivation for adoption [29].

2.4.7. Fusion Variability

- Fusion results can vary depending on the imaging protocol and equipment, complicating standardization [30, 31].

2.5. CNN-Based Multi-modal Medical Image Fusion: Step-By-Step Procedure

2.5.1. Step 1: Acquisition of Multi-Modal Medical Images

Once you have multiple medical images of the same subject captured from different imaging modalities like CT, MRI, PET, and Ultrasound, *etc.* Each modality provides distinct information.

2.5.2. Step 2: Preprocessing of Input Images

Next, the images are preprocessed to conform to the CNN model compatibility. It usually encompasses resizing of images, normalizing of pixel values and aligning the images by correcting some small spatial variations. To help the model better generalize the values, normalization (for

instance, scaling the pixel values into a 0–1 range) is used, which puts all values of an additional modality within a standardized range.

2.5.3. Step 3: Feature Extraction Using CNN Layers

Each preprocessed input image is passed through some successive CNN layers, just like convolutional layers that apply a filter to find edges, textures or some pattern in the image. Use of pooling layers to decrease the number of spatial dimensions in feature maps, keeping essential information, and restricting the computational load. The CNN learns to keep both anatomical and functional features specific to each modality, which then helps it distinguish between anatomical and functional features.

2.5.4. Step 4: Fusion of Extracted Features

Following feature extraction, the features obtained from each modality are integrated by means of a fusion strategy. CNNs typically allow for three types of fusion.

2.5.4.1. Pixel-Level Fusion

Captures detailed spatial information at the channel using the raw pixel values, which might increase the noise.

2.5.4.2. Feature-Level Fusion

Merges feature maps from different modalities while preserving abstract information and yields more informative and compact representations.

2.5.4.3. Decision-Level Fusion

Outcomes are combined after individual processing of each modality, a step specifically useful for high-level diagnostic tasks.

Depending on the focus to be applied in the fused image, common strategies include addition, averaging, maximum selection, or concatenation.

2.5.5. Step 5: Reconstruction of Fused Image

To reconstruct the fused feature map back in the image format, a deconvolution or upsampling layer is utilized. The goal of the reconstruction stage is to synthesize a unified image that maintains crucial features from each modality as a high-quality image with additional diagnostic value.

2.5.6. Step 6: Post-processing

Finally, post-processing steps were applied to further refine the fused image (*i.e.*, adjust contrast, use noise reduction filters, *etc.*) At this stage, sometimes the visibility of anatomical structures or functional details is enhanced, and the final output image is visually clear and diagnostically useful. The merits and demerits of CNN-based multi-modal medical image fusion are shown in Table 1.

2.6. Model Optimization, Adaptation, Future Research and Development

The model is fine-tuned to improve the CNN-based fusion process by optimizing parameters, trying various fusion strategies and employing advanced techniques such as transfer learning to make the model applicable for different datasets or imaging conditions. By optimization, the model is effective in various scenarios and can be gene-

ralized as applied to different patient cases and imaging equipment.

Table 1. Advantages and disadvantages of CNN-based multi-modal medical image fusion.

Advantages	Disadvantages
Automatic Feature Learning	High Data Requirements
High Detail Preservation	Computational Complexity
Scalability	Lack of Interpretability
Flexibility in Fusion Levels	Generalization Challenges
Improved Diagnostic Accuracy	Risk of Over-Smoothing

Future research may include further lightening of architectures to lessen computational demands, increasing the interpretation of models to increase acceptance from clinicians, and learning to adapt models to unseen imaging protocols. However, future advancements in this novel fusion model may lower the complexity of the CNNs to the point where they could be more feasibly incorporated into real-world healthcare settings, expanding the number of fused multimodal imaging scenarios that can be utilized for accurate and time-effective diagnosis and treatment. The general CNN-based multi-modal medical image fusion is shown in Fig. (2).

2.7. Algorithm: CNN-Based Multi-Modal Medical Image Fusion

The step-by-step process of fusing the source input images into a final fused image using a CNN. The CNN-based approach is explained in the sub-section.

2.7.1. Step 1. Input

CT_image β Input medical image 1, for example, CT image

MRI_image β Input medical image 2, for example, MRI image

2.7.2. Step 2. Preprocessing

CT_Image β Normalize (Resize(CT_Image, Size), Range = [0, 1])

MRI_Image β Normalize (Resize(MRI_Image, Size), Range = [0, 1])

2.7.3. Step 3. CNN Feature Extraction

Model β Define_CNN(Input_Shape = (Height, Width, Channels))

CT_Features β Model (CT_Image)

MRI_Features β Model (MRI_Image)

2.7.4. Step 4. Fusion Rule Application

For each Pixel (i, j):

Weight_CT β Compute_Weight(CT_Features(i, j))

Weight_MRI β Compute_Weight(MRI_Features(i, j))

Fused_Feature (i, j) β (Weight_CT * CT_Features(i, j)) + (Weight_MRI * MRI_Features (i, j))

2.7.5. Step 5. Post-Processing

Fused_Image β Reconstruct_Image(Fused_Feature)

Fused_Image β Apply_Filter (Fused_Image, Filter_Type = "Enhancement")

2.7.6. Step 6. Output

Display (Fused_Image)

Save (Fused_Image, Path = "Fused_Image.png")

2.7.7. Step 7. End

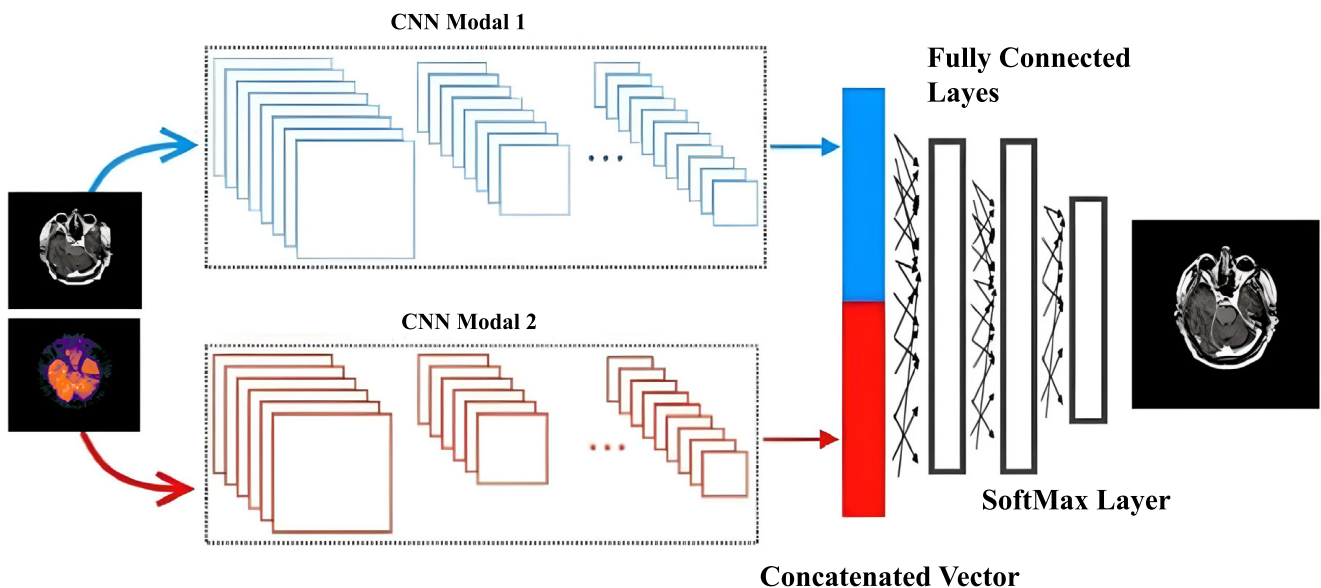


Fig. (2). General CNN-based multi-modal medical image fusion.

3. RESULTS

Figs. (3a, b and 4a, b) illustrate the datasets of CT, MRI, and PET images [32, 33]. The dataset used in this paper for experimental purposes is publicly available, as it is an open-source database. Figs. (3c-j and 4c-j) show the visual quality outcomes of the comparison of CNN with other prevalent fusion methods on two different datasets. Zooming into the images reveals that the results are good. The edges and corners are effectively preserved. Uniformity is maintained, and no information is lost. The CNN is effective in fusing the CT and MRI images without loss of any information. Additionally, the results of CNN are also compared with various standard methods as shown in Figs. (3c-j and 4c-j). The CNN is compared with conventional methods like Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT), Stationary Wavelet Transform (SWT), Wavelet Packet Decomposition (WPD), Multi-Singular Valued Dependency (MSVD), Non-Subsampled Contourlet Transform (NSCT), and Non-Subsampled Shearlet Transform (NSST), and the results are comparatively better than traditional methods in terms of visual and qualitative analysis. Upon thorough examination of the visual outcomes (Figs. 3c-j and 4c-j) and the parametric data (Tables 2 and 3), it can be concluded that the CNN

demonstrates superior performance relative to the other standard methods evaluated. The NSCT and NSST method also shows better results that look competitive in comparison to CNN-based results. The comparative performance of CNN is better than all the conventional methods as shown in Figs. (3c-j, 4c-j) and Tables 2 and 3. It can be concluded that the CNN-based multi-modal medical image fusion shows comparatively better fusion results than traditional methods.

The CNN-based method excels in spatial detail retention, feature preservation, and overall visual quality. Although there is a slight trade-off in structural similarity, it still produces visually appealing images with high contrast and detail. The NSCT and NSST methods show balanced performance across metrics, providing good structural similarity and quality. These methods are ideal for tasks requiring both feature retention and coherence. The PCA makes it suitable for applications focusing on feature-rich fusion, though it sacrifices some spatial correlation. The WPD and DWT methods are weaker in retaining structural and feature-related information, leading to fused images that may appear less detailed or visually coherent. The SWT method indicates potential issues with contrast and information richness in the resulting images.

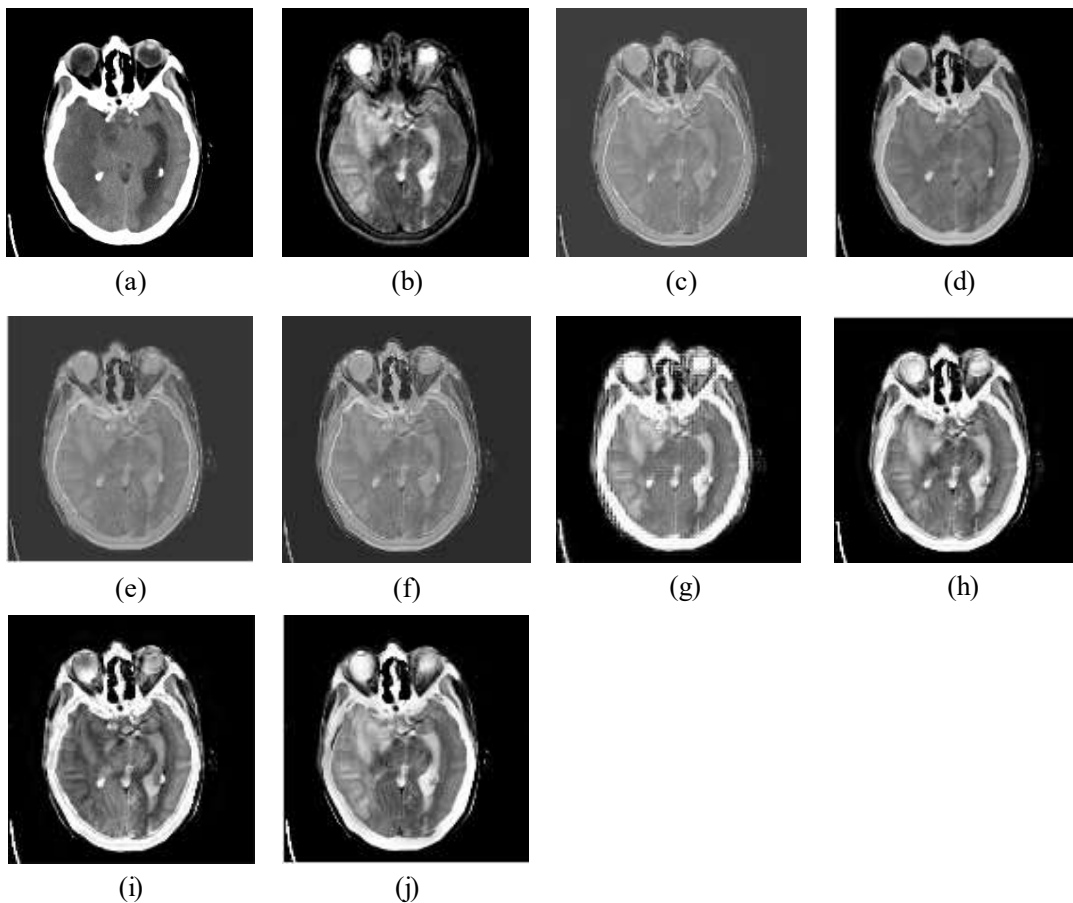


Fig. (3). (a) CT image, (b) MRI image, (c) MSVD, (d) PCA, (e) DWT, (f) SWT, (g) WPD, (h) NSCT, (i) NSST, (j) CNN.

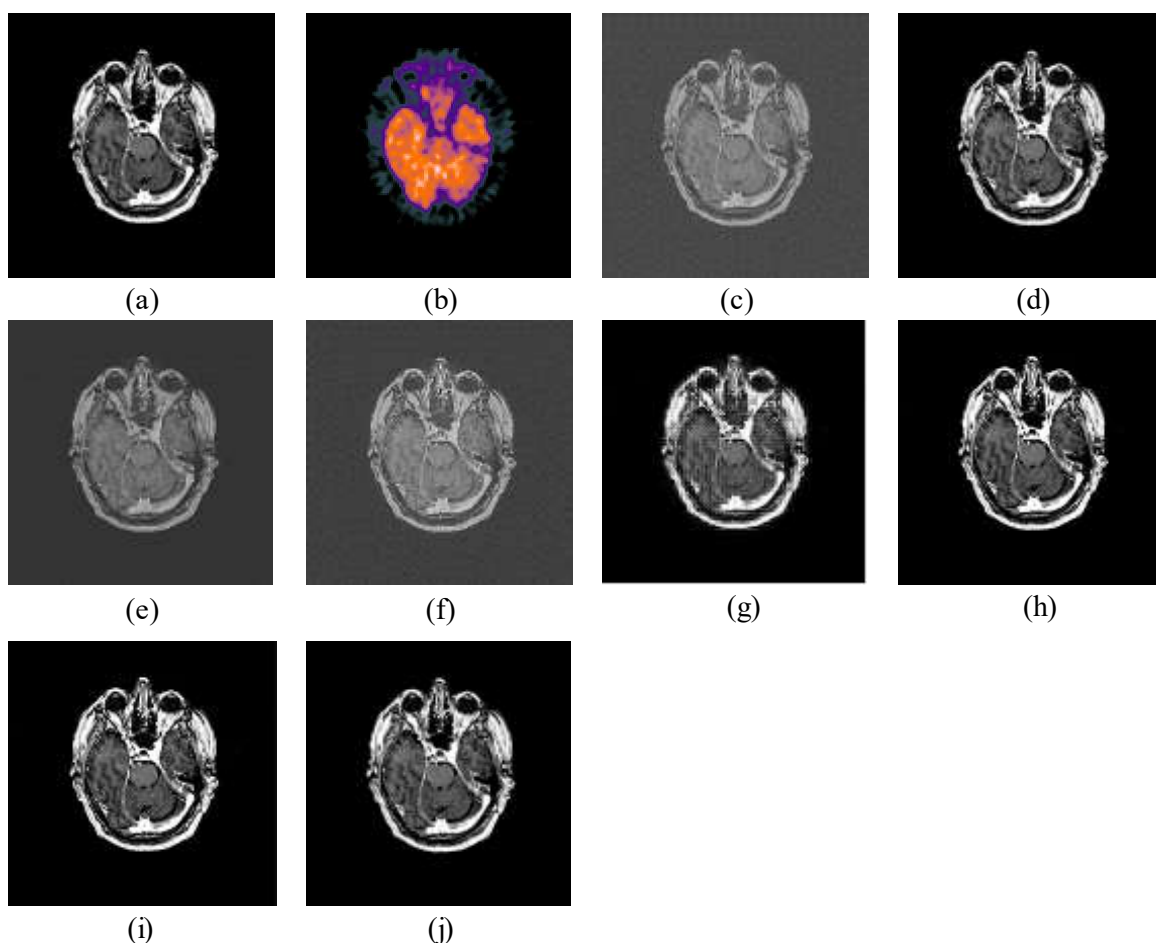


Fig. (4). (a) MRI image, (b) PET image, (c) MSVD, (d) PCA, (e) DWT, (f) SWT, (g) WPD, (h) NSCT, (i) NSST, (j) CNN.

Table 2. Average quantitative analysis as per SCD, FS, FMI, and FF.

Method	SCD	FS	FMI	FF
MSVD	1.38	0.98	0.70	1.22
WPD	1.28	0.92	0.71	1.45
PCA	1.32	0.93	1.11	1.51
DWT	1.65	0.98	0.90	0.98
SWT	1.39	0.98	0.91	1.35
NSCT	1.45	0.99	0.98	1.71
NSST	1.43	0.91	0.95	1.74
CNN	1.70	0.86	1.21	1.79

The CNN-based method excels in all metrics, producing visually rich, high-contrast, and well-balanced fused images. It is the best choice when visually superior fused images are required, such as medical imaging or surveillance. The NSCT and NSST methods provide balanced performance across all metrics, making them suitable for applications requiring good visual quality with slightly lower computational demands compared to CNN. The WPD method performs moderately well in entropy and fusion index, making it a viable choice for scenarios where high entropy

is prioritized over contrast. The PCA method shows reasonable performance but falls short in contrast and detail retention. The MSVD and DWT method shows poor performance, leading to visually dull images with low contrast, brightness, and information content.

Table 3. Average quantitative analysis as per SD, M, E, and FI.

Method	SD	M	E	FI
MSVD	40.11	19.52	0.90	20.17
WPD	60.70	30.65	3.34	31.56
PCA	56.55	26.57	0.98	28.03
DWT	38.60	27.50	1.31	22.47
SWT	40.90	19.52	1.38	20.60
NSCT	63.32	29.11	3.01	31.81
NSST	63.76	29.13	2.77	31.88
CNN	64.91	31.89	3.51	32.31

4. DISCUSSION

There is no reference image for the comparison. Therefore, objective method performance assessment metrics without a reference image are used for the overall com-

parison of M3IF-NSST-MI with other methods. The metrics used are Functional Mutual Information (FMI), Fusion Symmetry (FS), Mean (M), Fusion Factor (FF), Standard Deviation (SD), Entropy (E), Fused Index (FI), and Sum of Correlation Difference (SCD). Except for FS, all the metrics should have higher values for better results. Only FS should have a lower value for better results. Table 2 shows the quantitative analysis of different methods (MSVD, WPD, PCA, DWT, SWT, NSCT, NSST, and CNN) based on four metrics, SCD, FS, FMI, and FF. The key observations include the high scores of SCD and CNN. Whereas, NSCT scores the highest in terms of FS score, suggesting better feature similarity, while CNN has the lowest. CNN outperforms other methods with the highest FMI value, reflecting improved fusion metric performance. CNN also achieves the highest FF score, showcasing its strong fusion capabilities. Based on this data, CNN-based fusion methods shine in overall performance metrics, particularly SCD, FMI, and FF.

Table 3 evaluates different methods (MSVD, WPD, PCA, DWT, SWT, NSCT, NSST, and CNN) based on four metrics, SD, M, E, and FI. The key insights from these evaluations demonstrate that CNN achieves the highest SD score, which indicates the best variability in the fused image. CNN outperforms other methods with the highest mean value, suggesting better brightness retention in the fused image. CNN also achieves the highest E value, reflecting the most information-rich fused image. CNN secures the highest FI score, demonstrating the most efficient and balanced fusion

performance. CNN leads in all metrics (SD, M, E, and FI), indicating superior performance in fusion compared to other methods.

The graphical representation of the data presented in Tables 2 and 3 is illustrated in Figs. (5 and 6), respectively, using a spider chart. The findings presented in Tables 2 and 3, as well as in Figs. (5 and 6), represent average outcomes derived from experiments conducted on over 100 datasets. In Fig. (5), the CNN method covers the largest area on the chart, highlighting its superior performance in SCD (1.7), FMI (1.21), and FF (1.79). It shows a trade-off in FS (0.86), which is lower compared to other methods, but is compensated by its high scores in other metrics. The NSCT and NSST methods show balanced performance with slightly lower scores than CNN in FF and SCD, but higher scores in FS (0.99 for NSCT and 0.91 for NSST). The PCA method performs moderately well in FMI (1.11) and FF (1.51) but lags in SCD and FS. The WPD and SWT methods have a smaller area compared to CNN, indicating weaker overall performance. For example, WPD has a low FMI (0.71), while SWT struggles in FF (1.35). The MSVD and DWT methods show subpar performance in metrics like FMI and FF, with a relatively smaller contribution to the overall chart area.

In Fig. (6), the CNN method covers the largest area in the chart, clearly outperforming all other methods. It has the highest SD score, indicating superior contrast and detail retention.

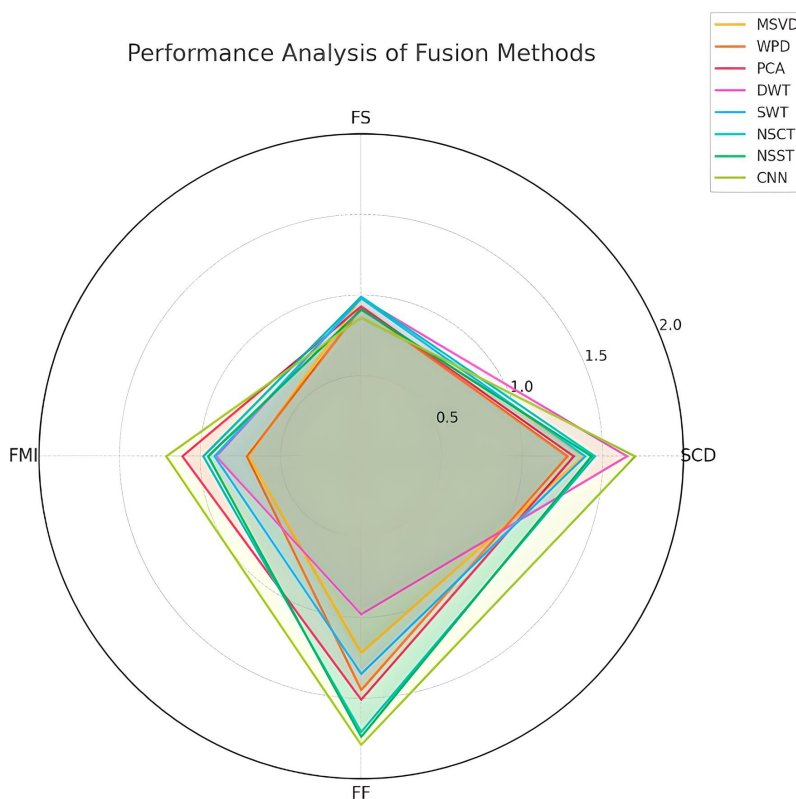


Fig. (5). Spider chart analysis of various metrics (SCD, FS, FMI, and FF).

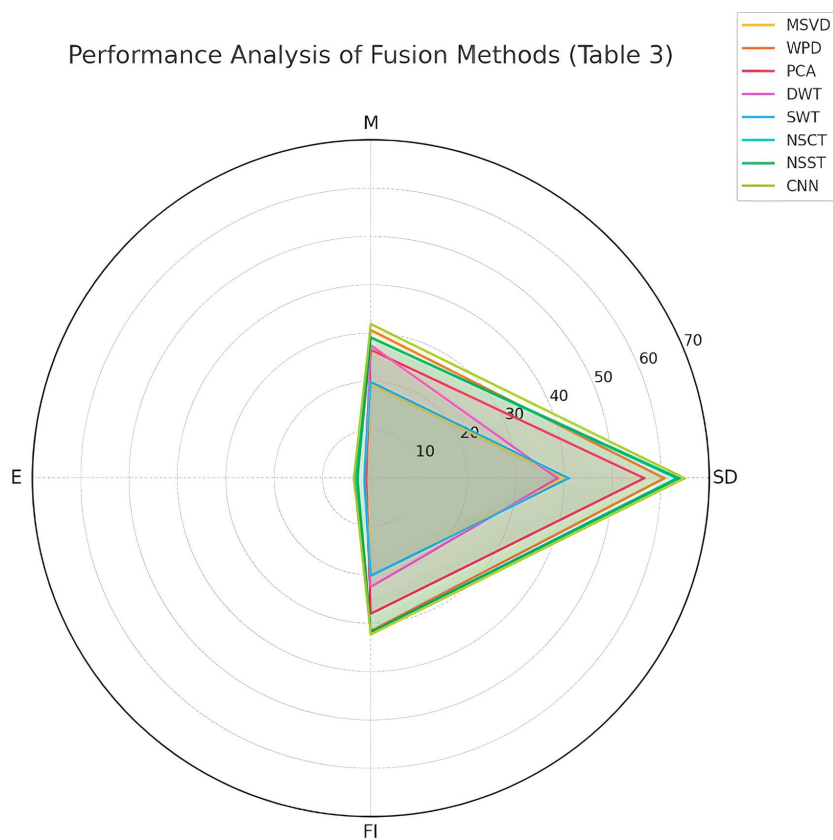


Fig. (6). Spider chart analysis of various metrics (SD, M, E and FI).

As well as the highest M score, reflecting better brightness preservation. Additionally, it obtains the highest E score, showcasing superior information retention, and overall has the best FI value, solidifying its dominance. The NSCT and NSST methods perform well and are closely aligned with CNN in most metrics, with their SD scores of 63.32 and 63.76, respectively, nearly as high as CNN. Their FI values are 31.81 and 31.88, respectively, competitive with CNN, demonstrating effective overall fusion performance. Whereas, their E scores are slightly lower than CNN but still above most other methods. The PCA method performs moderately, with good scores in M (26.57) and FI (28.03),

but lags in E (0.98) and SD (56.55). The WPD method is the second-best in FI (31.56) and performs decently in M (30.65) but falls behind in SD (60.7) and E (3.34). The MSVD and DWT methods underperform across all metrics, with particularly low FI and E, indicating limited utility for high-quality fusion tasks.

Since the fusion results difference are not visible in the Figs. (3 and 4) with the naked eye. Fig. (7) is introduced in this paper, which shows a zoomed-in view of Fig. (3), highlighting the superior performance of CNN-based fusion methods.

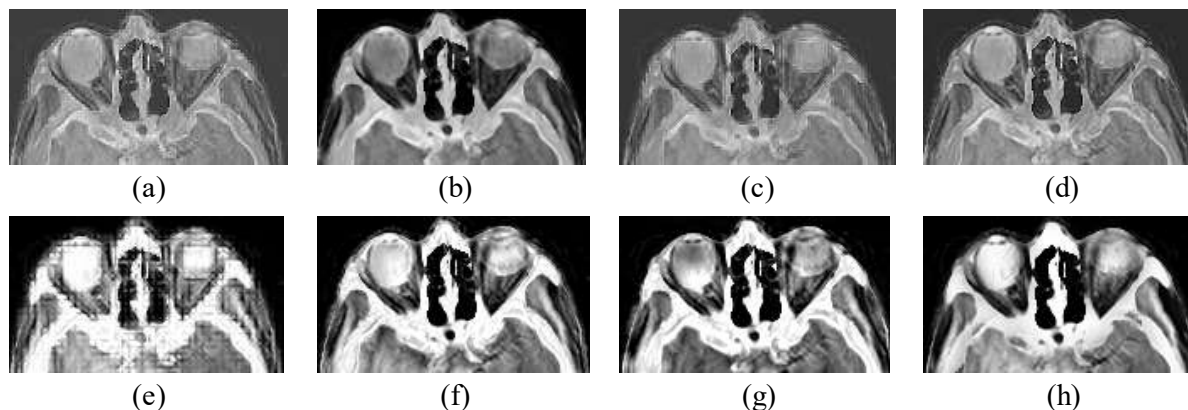


Fig. (7). Zooming fusion results (a) MSVD, (b) PCA, (c) DWT, (d) SWT, (e) WPD, (f) NSCT, (g) NSST, (h) CNN.

The image is visibly smooth and maintained in CNN-based fusion results in comparison to other methods. Although NSCT and NSST-based fusion results also show decent results in terms of tissue and other fine details preservation but overall clarity can be seen in the CNN-based method.

CONCLUSION

This paper examines the significance of CNN-based multimodal medical image fusion, combining data from modalities like CT, PET, and MRI into a single, informative image for enhanced diagnosis and treatment planning. CNN-based fusion integrates complementary anatomical and functional details, enabling automated feature extraction, reduced manual input, and greater diagnostic precision. Key benefits include lower healthcare costs and improved efficiency, particularly valuable in oncology, neurology, and cardiovascular imaging. Challenges include high computational demands, data requirements, and limited interpretability. Solutions such as lightweight architectures, transfer learning, and explainable AI are discussed, highlighting CNN-based fusion's potential to improve patient outcomes and advance medical imaging. In the coming years, advancements in CNN-based multimodal image fusion are poised to transform medical imaging by integrating deep learning innovations with efficient network architectures and enhanced interpretability. Reducing the computational demands of CNN models holds promise for making CNN-based fusion feasible in clinical settings. Data scarcity, a significant challenge, can be addressed through transfer learning and domain adaptation, enabling models trained on specific datasets to generalize effectively across different imaging conditions. Furthermore, incorporating Explainable AI (XAI) into CNN-based fusion frameworks could significantly improve model interpretability, helping clinicians understand both the fusion processes and the resulting outputs. In diagnostic applications, where accuracy is paramount, transparency becomes especially critical. CNN-driven multimodal image fusion offers tremendous potential for enhancing healthcare by combining diverse imaging modalities to improve diagnosis, treatment planning, and patient outcomes. Although challenges remain, current research and technological advancements are expected to mitigate these limitations, pushing the boundaries of medical imaging through CNN-based fusion.

AUTHORS' CONTRIBUTIONS

S.B, P.S., D.G., V.R., M.D.: Methodology.

LIST OF ABBREVIATIONS

CT	=	Computed Tomography
PET	=	Positron Emission Tomography
MRI	=	Magnetic Resonance Imaging
CNN	=	Convolutional Neural Network

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data and supportive information are available within the article.

FUNDING

None.

CONFLICT OF INTEREST

Dr. Vinayakumar Ravi is the Associate Editorial Board Member of the journal The Open Bioinformatics Journal.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

- [1] Hermessi H, Mourali O, Zagrouba E. Convolutional neural network-based multimodal image fusion via similarity learning in the shearlet domain. *Neural Comput Appl* 2018; 30(7): 2029-45. <http://dx.doi.org/10.1007/s00521-018-3441-1>
- [2] Almasri MM, Alajlan AM. Artificial intelligence-based multimodal medical image fusion using hybrid S2 optimal CNN. *Electronics* 2022; 11(14): 2124. <http://dx.doi.org/10.3390/electronics11142124>
- [3] Singh S, Anand RS. Multimodal medical image fusion using hybrid layer decomposition with CNN-based feature mapping and structural clustering. *IEEE Trans Instrum Meas* 2020; 69(6): 3855-65. <http://dx.doi.org/10.1109/TIM.2019.2933341>
- [4] Wang L, Zhang J, Liu Y, Mi J, Zhang J. Multimodal medical image fusion based on Gabor representation combination of multi-CNN and fuzzy neural network. *IEEE Access* 2021; 9: 67634-47. <http://dx.doi.org/10.1109/ACCESS.2021.3075953>
- [5] Liu S, Liu H, John V, Liu Z, Blasch E. Enhanced situation awareness through CNN-based deep multimodal image fusion. *Opt Eng* 2020; 59(5): 1. <http://dx.doi.org/10.1117/1.OE.59.5.053103>
- [6] Li Y, Zhao J, Lv Z, Pan Z. Multimodal medical supervised image fusion method by CNN. *Front Neurosci* 2021; 15:638976 <http://dx.doi.org/10.3389/fnins.2021.638976> PMID: 34149344
- [7] Azam MA, Khan KB, Salahuddin S, *et al.* A review on multimodal medical image fusion: Compensious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput Biol Med* 2022; 144:105253 <http://dx.doi.org/10.1016/j.compbiomed.2022.105253> PMID: 35245696
- [8] Hermessi H, Mourali O, Zagrouba E. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing* 2021; 183:108036 <http://dx.doi.org/10.1016/j.sigpro.2021.108036>
- [9] Tan W, Tiwari P, Pandey HM, Moreira C, Jaiswal AK. Multimodal medical image fusion algorithm in the era of big data. *Neural Comput Appl* 2020; 1-21. <http://dx.doi.org/10.1007/s00521-020-05173-2>
- [10] Guo Z, Li X, Huang H, Guo N, Li Q. Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington, DC, USA, 04-07 April 2018, pp. 903-907
- [11] Maneesha P, Singh T, Nayar R, Kumar S. Multi modal medical image fusion using convolution neural network. 2019 Third International Conference on Inventive Systems and Control (ICISC). Coimbatore, India, 10-11 January 2019, pp. 351-357 <http://dx.doi.org/10.1109/ICISC44355.2019.9036373>

- [12] Liu S, Yin L, Miao S, Ma J, Cong S, Hu S. Multimodal medical image fusion using rolling guidance filter with CNN and nuclear norm minimization. *Curr Med Imaging Rev* 2021; 16(10): 1243-58. <http://dx.doi.org/10.2174/1573405616999200817103920> PMID: 32807062
- [13] Song L, Liu J, Qian B, et al. A deep multi-modal CNN for multi-instance multi-label image classification. *IEEE Trans Image Process* 2018; 27(12): 6025-38. <http://dx.doi.org/10.1109/TIP.2018.2864920> PMID: 30106729
- [14] Goyal S, Singh V, Rani A, Yadav N. Multimodal image fusion and denoising in NSCT domain using CNN and FOTGV. *Biomed Signal Process Control* 2022; 71103214 <http://dx.doi.org/10.1016/j.bspc.2021.103214>
- [15] Castro FM, Marín-Jiménez MJ, Guil N, Pérez de la Blanca N. Multimodal feature fusion for CNN-based gait recognition: An empirical comparison. *Neural Comput Appl* 2020; 32(17): 14173-93. <http://dx.doi.org/10.1007/s00521-020-04811-z>
- [16] Joze HRV, Shaban A, Iuzzolino ML, Koishida K. MMTM: Multimodal transfer module for CNN fusion. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 13289-13299
- [17] Rajalingam B, Priya R. Multimodal medical image fusion based on deep learning neural network for clinical treatment analysis. *Int J Chemtech Res* 2018; 11(06): 160-76.
- [18] Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf Fusion* 2020; 54: 99-118. <http://dx.doi.org/10.1016/j.inffus.2019.07.011>
- [19] Song J, Zheng J, Li P, Lu X, Zhu G, Shen P. An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Front Digit Health* 2021; 3637386 <http://dx.doi.org/10.3389/fgdth.2021.637386> PMID: 34713109
- [20] Wang D, Mao K, Ng GW. Convolutional neural networks and multimodal fusion for text aided image classification. 2017 20th International Conference on Information Fusion (Fusion). Xi'an, China, 10-13 July 2017, pp. 1-7 <http://dx.doi.org/10.23919/ICIF.2017.8009768>
- [21] Guo K, Li X, Hu X, Liu J, Fan T. Hahn-PCNN-CNN: An end-to-end multi-modal brain medical image fusion framework useful for clinical diagnosis. *BMC Med Imaging* 2021; 21(1): 111. <http://dx.doi.org/10.1186/s12880-021-00642-z> PMID: 34261452
- [22] Muzammil SR, Maqsood S, Haider S, Damaševičius R. CSID: A novel multimodal image fusion algorithm for enhanced clinical diagnosis. *Diagnostics* 2020; 10(11): 904. <http://dx.doi.org/10.3390/diagnostics10110904> PMID: 33167376
- [23] Wang L, Zhang H, Yang J. Finger multimodal features fusion and recognition based on CNN. 2019 IEEE Symposium Series on Computational Intelligence (SSCI). Xiamen, China, 06-09 December 2019, pp. 3183-3188 <http://dx.doi.org/10.1109/SSCI44817.2019.9003093>
- [24] Deng X, Dragotti PL. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans Pattern Anal Mach Intell* 2021; 43(10): 3333-48. <http://dx.doi.org/10.1109/TPAMI.2020.2984244> PMID: 32248098
- [25] Rajalingam B, Al-Turjman F, Santhoshkumar R, Rajesh M. Intelligent multimodal medical image fusion with deep guided filtering. *Multimedia Syst* 2022; 28(4): 1449-63. <http://dx.doi.org/10.1007/s00530-020-00706-0>
- [26] Münzner S, Schmidt P, Reiss A, Hanselmann M, Stiefelhofen R, Dürichen R. CNN-based sensor fusion techniques for multimodal human activity recognition. *Proceedings of the 2017 ACM international symposium on wearable computers*. Maui, Hawaii, 2017-, pp. 158-165 <http://dx.doi.org/10.1145/3123021.3123046>
- [27] Kumar V, Joshi K, Kumar R, et al. Multi modalities medical image fusion using Deep Learning and metaverse technology: Healthcare 4.0 A futuristic approach. *Biomed Pharmacol J* 2023; 16(4): 1949-59. <http://dx.doi.org/10.13005/bpj/2772>
- [28] Kumar V, Joshi K, Kanti P, Reshi JS, Rawat G, Kumar A. Brain tumor diagnosis using image fusion and deep learning. 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). Erode, India, 23-25 March 2023, pp. 1658-1662 <http://dx.doi.org/10.1109/ICSCDS56580.2023.10104937>
- [29] Diwakar M, Tripathi A, Joshi K, et al. A comparative review: Medical image fusion using SWT and DWT. *Mater Today Proc* 2021; 37: 3411-6. <http://dx.doi.org/10.1016/j.matpr.2020.09.278>
- [30] Joshi K, Kumar M, Tripathi A, Kumar A, Sehgal J, Barthwal A. Latest trends in multi-modality medical image fusion: A generic review. *Rising Threats in Expert Applications and Solutions Lecture Notes in Networks and Systems* Springer, Singapore, vol 434, pp. 663-671.
- [31] Joshi K, Kumar V, Sundaresan V, et al. Intelligent fusion approach for MRI and CT imaging using CNN with wavelet transform approach. 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES). Chickballapur, India, 28-29 December 2022, pp. 1-6 <http://dx.doi.org/10.1109/ICKES56523.2022.10060322>
- [32] Zero learning fast medical image fusion. 2024. Available from: <https://github.com/bsun0802/Zero-Learning-Fast-Medical-Image-Fusion/tree/master/images/MRI-PET>
- [33] Multimodal image fusion to detect brain tumors. 2024. Available from: <https://github.com/ashna111/multimodal-image-fusion-to-detect-brain-tumors/tree/master/dataset>