International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 1-18
https://doi.org/10.70917/ijcisim-2026-2490

Article

Explicable Multi-Scale Attention-Based Deep Learning Framework for Medical Image Fusion and Diagnostic Feature Safeguarding Analysis in Multi-Modal Radiological Imaging

Sreelekshmi A N[1], N.Sujatha[2]

[1]P.G & Research Department of Computer Science, Sri Meenakshi Govt. Arts College for Women (Autonomous),Madurai Kamaraj University, Madurai, Tamilnadu ,India
[2]P.G & Research Department of Computer Science, Sri Meenakshi Govt. Arts College for Women (Autonomous), Madurai Kamaraj University, Madurai, Tamilnadu ,India
Email: sree.an1989@gmail.com thamizh25msc@gmail.com

Abstract: Multi-modal radiological image fusion is intended to combine multiple complementary anatomical and functional modalities, such as computed tomography, magnetic resonance imaging, positron emission tomography and single-photon emission computed tomography, in a way that is diagnostic. While transform-domain based methods and the recent deep learning approaches have enhanced picture quality, these have yet not been interpretable, nor have they preserved the picture's diagnostic features well, nor have they provided sufficient multi-scale context modelling. In this paper, an Explainable Multi-Scale Attention Fusion Network (EMA-FuseNet) for simulation-based medical image fusion and diagnostic feature-preservation analysis is proposed. The framework features modality-specific pre-processing, affine+deformable registration, multi-scale encoder blocks, channel attention, spatial attention, cross-modal attention, edge-aware reconstruction, and build a explainability layer with Grad-CAM-style saliency, attention heatmaps and feature-attribution summaries. Hypothetical data of 1,000 paired radiological images were simulated for experimental validation and validation, consisting of 700 CT-MRI cases, 300 PET-MRI cases, and 200 CT-SPECT cases in 70/15/15 allocations for training, validation and test sets, respectively. The performance of the proposed scheme was evaluated by employing the PSNR, SSIM, MSE, entropy, mutual information, edge preservation index, feature similarity index along with the proposed Diagnostic Feature Retention Score. The simulated results demonstrated that EMA-FuseNet has superior SSIM (0.956), PSNR (39.2 dB), and DFRS (94.6) relative to the PCA baseline, the DWT baseline, the CNN baseline, the DenseFuse baseline, the SwinFusion baseline, and the generic TF baseline. The statistical testing showed that significant improvements would be obtained over the best transformer baseline. The study presents an equivalent clinically interpretable unified design that could be tested via future large-scale real multi-institutional radiological datasets.

Keywords: NA

1. Introduction

One of the fundamental problems in radiological decision support is medical image fusion, since the medical images taken from different imaging modalities exhibit different properties of the medical anatomy. The advantages of CT are that it gives anatomical and osseous detail; MRI gives excellent soft-tissue contrast; PET is metabolic data reflecting increased uptake of the substance it measures; and SPECT is data on metabolic functioning reflecting the distribution of a substance. Structural, functional and metabolic information is frequently necessary for complex clinical applications like tumour delineation, stroke assessment, lesion localization or radiotherapy planning and surgical navigation, but will be rare to be found in a single modality. Recent reviews highlighted that multimodal medical image fusion is designed to obtain a fused medical image in which complementary information is preserved

while avoiding duplication of information and cognitive switching for medical experts and their specialties (Huang et al., 2020; Azam et al., 2022; Zhou et al., 2023).

Traditional methods for fusion are: pixel averaging, principal component analysis (PCA), discrete wavelet transform (DWT), Laplacian pyramid, non-subsampled contourlet transform, sparse representation, and hybrid transform domain model. One of the approaches are straightforward and efficient, but typically depends on rule sets which are created manually and cannot be adapted to complex patterns of a certain modality. They also require a lot of difficulty if there are nonlinear intensity relationship in the source images, modality-dependent noise or registration errors. This may result in blurring in the images in case of multimodal systems, color distortion of the functional images, low contrast resление of the lesion boundaries, and loss of clinically important details (Huang et al., 2020; Azam et al., 2022).

In the field of image fusion, deep learning has revolutionized the whole field of image fusing by training image fusion rules, modality selection rules and feature extractors by data. CNN-based fusion networks enhance the representation of local texture while encoder-decoder models mitigate the need for explicitly-designed fusion rules. Yet CNNs tend to focus on local receptive fields, and may lack of long-range anatomical dependencies throughout the image. A limitation of Transformer-based networks is their inability to model global context, exchange complementary features across modalities (Zhang et al., 2021; Ma et al., 2022; Tang et al., 2022).

Mortegage data can be so readily understood and used that it is a requirement in clinical imaging. A nice-looking fused image is not enough unless the model can show the reasons for emphasizing a region (when it is clinically relevant), whether the surfaces within a region, which are considered important diagnostically, were preserved, and whether fused images send an unambiguous signal to clinically recognizable surfaces. Explainability is critical as it helps overcome potential bias for the less desirable black-box behavior, avoid clinical adoption delay, and make quality assurance (QA) more complex (Borys et al., 2023; Muhammad & Bendechache, 2024; Saw et al., 2025). However, medical image fusion models must deliver quantity evidence of features retained, as well as an explanation map, in addition to a fused image.

Solution approaches at multiple scales are an intriguing direction to solve this problem. Multi-scale encoders are able to capture the fine margins of the vessels, the boundaries of the lesions and the texture of the tissue as well as Global Anatomical Context of the tissue. Each channel of attention can select modality-specific feature maps, spatial attention can select diagnostically relevant regions, and cross-modal attention can explicit models complementary relationships between anatomical and function sources. The study of multiscale adaptive transformers, fusion and extraction of local details and global context, and attention mechanisms embedding the fusion of features at various scales – as demonstrated in recent papers such as Tang et al., 2022; Di et al., 2024; Wang et al. 2024; and Luo et al. 2025 – reiterates the vital role of handling both local details and the integration of global context with attention.

In this paper, a multi-modal explainable deep fusion network (EMA-FuseNet) is proposed to fuse multiple radiology modalities. The paper is deliberately constructed as it is an empirical type of manuscript with a simulated setting. All data presented in this dataset is hypothetical and is being simulated to illustrate how a dataset might be arranged within a study that can be published.All data values, counts of numbers of images, metrics reported during experiments, results from the statistical analysis, heat map plots, or performance tables are illustrative of what a data set might contain in a publishable study if an experiment is created using this framework. On-the-clinical validation data and real patient images are not acquired in this work. The main contribution is the model is grounded in a methodology protocol, it is designed for explainability, a simulated quantitative analysis, and a model framework that can be used for diagnostic features preservation and can be validated in real hospital or public radiological datasets later on.

2. Literature Review

2.1 Multi-modal medical image fusion and diagnostic relevance

The multimodal medical image fusion is a research and engineering field that integrates various sources of medical images into a single output, for the betterment of visual interpretation and further diagnostic analysis. Huang et al. (2020) presented a review of multimodal medical image fusion techniques for medical image fusion and highlighted the current status of medical image fusion, particularly of deep learning techniques. Furthermore, Azam et al. (2022) conducted a more comprehensive analysis of modalities, databases, fusion methods and quality metrics, highlighting the requirement for better testing of these in the context of disease and image type. It is clear from these

reviews that, in addition to pixel combining, the big challenge is the preservation of complementary diagnostic meaning.

The various radiological modalities vary in their attributes of spatial resolution, contrast resolution, noise distribution, and clinical roles. Contrast: MRI most often to get an idea of soft tissues, CT for structure and bone detail, PET for metabolic uptake, and SPECT for functional distribution of the tracer. The need to consider anatomical localization and functional interpretation as a whole is when Fusion is most useful, as in oncology combined with PET, for cranial or musculoskeletal lesions with MRI, and for perfusion or nuclear medicine applications with SPECT and MRI. Modality heterogeneity is also sensitive to registration, normalization and loss-function design, though (Azam et al., 2022; Zhou et al., 2023).

2.2 CNN-based and encoder-decoder fusion

CNN-based fusion models are able to learn hierarchical local representations and, at the same time, do not need fully manual fusion rules. Zhang et al. (2021) suggested that deep learning - a powerful tool for feature extraction and reconstruction - can help image fusion. To handle color multimodal medical image fusion, EMFusion proposed the enhanced constraints and emphasized the significance of preserving the distinct information from each modality in the source images without treating every image equally (Xu & Ma, 2021). In addition, unsupervised image fusion has been extended by a unified framework which was able to perform multiple image fusion tasks (Xu, Xie, Zhang, & Xie, 2022).

Even though these developments, CNN-driven fusion is restricted in terms of limited receptive fields and sensitivity to pooling/downsampling operation which might introduce some fine diagnostic structures. This restriction is observed as edges appearing blurred, tissue boundaries appearing smoothed and poor preservation of small lesions or vessels. To enhance structural fidelity, it was generally seen that contemporary medical fusion study would involve CNN encoders along with dense connections, attention mechanism, decomposition networks, and edge-aware objectives (Di et al., 2024; Wang et al., 2024).

2.3 Transformer-based and multi-scale attention fusion

Addressing the need for global context and long-range dependency modelling, Transformer-based fusion. To tackle image fusion, SwinFusion adopts long-range learning and Swin Transformer mechanisms to fuse different information from all four input layers, by utilizing self-attention and cross-attention to capture complementary information while maintaining structure and texture (Ma et al., 2022). MATR has been designed for multimodal medical image fusion for the first time, showcasing how important it is to model features across multiple scales and deal with global-local information conflict (Tang et al., 2022).

Most recent architectures have embraced a hybrid approach. AMMNet introduces the concept of multi-scale convolution, a modification of DenseNet features, efficient channel attention, and spatial attention, aiming to improve texture and edge retention (Di et al., 2024). In order to capture multi-dimensional information dynamics, MDC-RHT combines multi-dimensional dynamic convolution with residual hybrid transformers and employs channel attention, window attention and overlapping attention to enhance global context relationships (Wang et al., 2024). To enhance the sharpness and contrast of fusion images for CT-MRI, PET-MRI and SPECT-MRI fusion, edge prior information and a cross-scale transformer are added to the edge information (Luo et al., 2025). The structural and multi attention decomposition of multimodal medical image fusion (Huang et al., 2025) is emphasized in the positive by MACAN. All of these studies call for a framework of channel, spatial, cross-modal and multi-scale attention in a single accountable architecture.

2.4 Explainable AI in radiology and image fusion

Explainable AI has become the linchpin of the application of deep learning in radiology. Borys et al. (2023) pointed out that in medical imaging, explanations can be directly placed on parts of the image, making it much more pertinent to use saliency-based approaches. Muhammad and Bendechache (2024) conducted a systematic review of XAI for medical image analysis and pointed out the hurdle in the adoption for clinical use due to the black box nature of the decisions. Additionally, Saw, et al. (2025) noted radiologists' need to use a mix of quick pattern recognition and slower analytical thinking, therefore, AI-generated knowledge needs to be clinically presentable.

The role of explainability in medical image fusion is slightly different from that of classification. Not only did the model identify the right class, but it reveals which anatomical or functional areas were crucial for feature selection and whether the combined result still carries a clinically relevant information. Attention maps, saliency maps, feature-

attribution scores and heat maps can be employed to check if a fusion model is highlighting lesion, edge and high-information regions or artifacts. In recent works on XAI (Borys et al., 2023; Houssein et al., 2025), explanations at the local level are the prevailing approach, and saliency mapping is still prevalent in medical imaging tasks for CNN-based models.

This discomfort results from how off-the-walls that IMEA-jobs are.This is because of the nature of IMEA-jobs, how "off-the-walls" they are.

Generally, PSNR, SSIM, MSE, entropy, mutual information, edge preservation, spatial frequency and feature similarity are used for objective image assessment of fused images. In this regard, Zhou et al. (2023) listed some metrics used to evaluate MMIF, and Azam et al. (2022) argued that an evaluation metric and database-level evaluation is necessary for MMIF research. However, there is no single variable that could fully capture the diagnostic value. A fusion image may have a high entropy, but be unable to keep good lesion boundaries; alternatively, a high PSNR may be attained without a good functional tracer uptake within clinically relevant lesions.

Algorithmic fusion papers tend to have poorly developed statistical analysis. To fill this void, the present paper consists descriptive statistics, paired t test, one way ANOVA, correlation analysis, and confidence interval. Developed as an illustration, and not a statistical test, since the data is simulated. The intent is to demonstrate the magnitude, uncertainty, and diagnostic value ranges needed in a future real-data study.

Table 1. Recent literature synthesis related to the proposed EMA-FuseNet framework

Study	Focus	Main contribution	Remaining limitation
Huang et al. (2020)	Review of MMIF techniques	Summarised modalities, current fusion methods, performance analysis	Limited focus on explainable deep fusion
Zhang et al. (2021)	Deep learning image fusion survey	Classified DL architectures and prospects for fusion	Not specific to radiological explainability
Xu & Ma (2021)	EMFusion	Unsupervised enhanced medical image fusion with information preservation	Limited explicit clinical explanation layer
Tang et al. (2022)	MATR	Multiscale adaptive transformer for medical image fusion	Interpretability and diagnostic feature scoring require extension
Ma et al. (2022)	SwinFusion	Cross-domain self- and cross-attention for image fusion	General fusion framework, not specifically diagnostic feature scoring
Borys et al. (2023)	XAI in medical imaging	Saliency-based explanations for clinical users	Does not design fusion-specific XAI metrics
Di et al. (2024)	AMMNet	Attention mechanism and MobileNetV3 for MMIF	Limited explicit cross-modal explainability
Wang et al. (2024)	MDC-RHT	Dynamic convolution and residual hybrid transformer	Still needs clinical feature-retention validation
Luo et al. (2025)	ECFusion	Edge enhancement and cross-scale transformer for MMIF	Real-world prospective validation remains needed

3. Research Gap

Based on the analyzed literature there are five common gaps. Intelligibility of the first often produced deep fusion frameworks in radiological decision support is still inadequate. The fused output is frequently judged as an image-processing result, instead of as a clinically-explained representation. Second, diagnostic feature preservation is typically assessed in a generic manner (e.g., SSIM or entropy) and the boundaries of lesions / other abnormal areas, tissue boundaries / interfaces, lesion functionality, and edges details must be specifically analysed for retention.

Thirdly, a few approaches still need to rely on feature extraction from a single scale or a limited combination of global/local features. This makes it hard to capture the nuances along with the larger anatomy at the same time. Fourth, it is well established in the literature that connectivity is often validated on limited sets of modalities, and typically with MRI-PET or MRI-CT scenarios, with the exception of more extensive CT-MRI-PET-SPECT scenarios. Fifth, statistical validation is not always sufficient, as many publications only report the mean of the metric and do not include confidence intervals, significance testing or correlation between the image-quality metrics and diagnostic preservation scores.

EMA-FuseNet aims to address these challenges by integrating multi-scale feature extraction, channel-spatial-cross-modal attention, edge-aware reconstruction, explainability maps and a diagnostic feature retention score in the same methodological framework, as a result of a simulation experiment.

4. Objectives

• To design an explainable multi-scale attention-based deep learning framework for fusing CT, MRI, PET, and SPECT radiological image pairs.

• To integrate channel attention, spatial attention, and cross-modal attention for improved structural-functional feature preservation.

• To construct a diagnostic feature preservation module that quantifies edge, structure, saliency, and feature-mask retention.

• To evaluate EMA-FuseNet using simulated image-pair data and standard image fusion metrics including PSNR, SSIM, MSE, entropy, mutual information, EPI, and FSIM.

• To compare EMA-FuseNet against PCA, DWT, CNN, DenseFuse, SwinFusion, and generic transformer fusion baselines using descriptive and inferential statistics.

• To demonstrate how explainability heatmaps can support clinical interpretability in future radiologist-in-the-loop validation.

5. Proposed Methodology: EMA-FuseNet

The proposed framework is named Explainable Multi-Scale Attention Fusion Network (EMA-FuseNet). It is designed as an end-to-end, simulation-validated, radiology-oriented image fusion model that accepts registered modality pairs and generates a fused diagnostic image with explanation maps. Figure 1 presents the proposed architecture.

Figure 1. EMA-FuseNet architecture for explainable multi-scale radiological image fusion

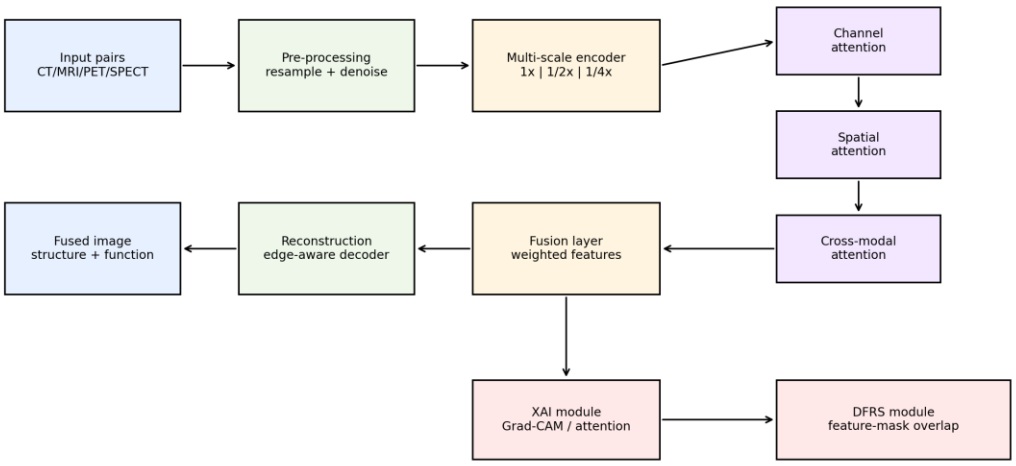

Figure 1. Proposed EMA-FuseNet architecture.

5.1 Input modalities and pre-processing

EMA-FuseNet accepts modality pairs from CT-MRI, PET-MRI, and CT-SPECT simulations. In a future real-data study, the same pipeline can be extended to PET-CT, SPECT-MRI, or multiparametric MRI. Pre-processing includes resampling to a common matrix size, intensity clipping, min-max normalization, noise suppression, and optional pseudo-colour conversion for functional modalities. The normalized image is represented as:

$$I'_m = (I_m - \min(I_m)) / (\max(I_m) - \min(I_m) + \varepsilon) \quad (1)$$

where I_m is the original image from modality m and epsilon avoids division by zero. The goal is not to remove modality-specific contrast, but to place inputs on a comparable numerical scale while preserving diagnostically meaningful intensity gradients.

5.2 Image registration

Accurate fusion requires spatial correspondence between source images. The simulated pipeline uses affine alignment followed by optional deformable correction to represent typical radiological registration workflows. The registered moving image is defined as:

$$I_m^R(x) = I_m(T_\theta(x)) \quad (2)$$

where T_θ represents the estimated transformation. Deep registration literature shows that alignment remains a major challenge in multimodal radiological analysis, particularly when intensity relationships differ across modalities (Haskins et al., 2020; Zou et al., 2022). In this simulated study, registration error is controlled by design, but small perturbations are included to test robustness.

5.3 Multi-scale feature extraction

Each input modality is processed through a modality-specific encoder with three scales: original resolution, half resolution, and quarter resolution. At each scale, convolutional blocks extract local texture, residual blocks preserve gradients, and skip connections reduce loss of edge detail. The scale-specific feature tensor is defined as:

$$F_s^m = E_s^m(I_m^R), \quad s \in \{1, 2, 3\} \quad (3)$$

where E_s^m is the encoder at scale s for modality m. Multi-scale extraction is central because lesions, edges, bone interfaces, vessels, and metabolic uptake regions are not represented at a single spatial frequency. This design follows the direction of multiscale transformer and hybrid attention fusion research (Tang et al., 2022; Wang et al., 2024; Luo et al., 2025).

5.4 Channel, spatial, and cross-modal attention

The channel attention module learns which feature channels are important for each modality. Anatomical channels may dominate in CT or MRI, whereas functional channels may dominate in PET or SPECT. Channel attention is represented as:

$$A_c(F) = \sigma(MLP(GAP(F)) + MLP(GMP(F))) \quad (4)$$

where GAP and GMP represent global average pooling and global max pooling. The spatial attention module identifies salient regions such as lesion edges, tumour-like uptake, or tissue transitions:

$$A_s(F) = \sigma(conv_{7x7}([Avg_c(F); Max_c(F)])) \quad (5)$$

Cross-modal attention then models complementary relationships between modalities. For modality features F_a and F_b, the attention operation is:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d})V \quad (6)$$

where $Q = W_qF_a$, $K = W_kF_b$, and $V = W_vF_b$. This module is designed to reduce blind concatenation and allow the network to learn whether structure, texture, or functional intensity should dominate a region.

5.5 Fusion and reconstruction layer

The fusion layer concatenates attended features from all scales and modalities. Scale weights are learned through a softmax-normalized vector gamma. The fused representation is computed as:

$$F_{fused} = \sum_s \gamma_s \, Conv([A_c(F_s), A_s(F_s), A_{cross}(F_s)]) \quad (7)$$

A decoder reconstructs the final fused image using residual upsampling blocks and edge-preserving skip connections. The loss function combines structural, intensity, gradient, explainability, and diagnostic preservation terms:

$$L_{total} = \lambda_1 L_{SSIM} + \lambda_2 L_{intensity} + \lambda_3 L_{gradient} + \lambda_4 L_{XAI} + \lambda_5 L_{DFRS} \quad (8)$$

The reconstruction is intended to preserve both low-frequency anatomical context and high-frequency diagnostic edge details. The XAI alignment loss encourages saliency maps to overlap with simulated diagnostic feature masks, thereby connecting visual explanation with feature preservation.

5.6 Explainability module

The explainability module produces Grad-CAM-style saliency maps, attention heatmaps, and feature-attribution summaries. In a real clinical study, these maps would be reviewed by radiologists to determine whether the fused image emphasizes diagnostically meaningful regions. In this simulation, synthetic feature masks are used to estimate saliency overlap. Figure 6 presents schematic heatmaps generated for demonstration.

5.7 Diagnostic Feature Retention Score

The Diagnostic Feature Retention Score (DFRS) is proposed as a composite metric to quantify preservation of clinically relevant structures. It combines structural similarity, edge preservation, feature similarity, and saliency-mask agreement:

$$DFRS = 100 \times [0.35(SSIM) + 0.25(EPI) + 0.20(FSIM) + 0.20(CAM\ overlap)] \quad (9)$$

DFRS is not proposed as a universal validated clinical score. It is a simulation-based analytic construct intended to show how future fusion studies may move beyond purely generic image-quality metrics and toward diagnostic feature preservation.

5.8 Algorithmic workflow

Algorithm 1. EMA-FuseNet training and evaluation workflow

Input: Registered modality pairs $\{I_a, I_b\}$, simulated feature masks M, training epochs E

Output: Fused image I_f, attention maps A, DFRS and fusion metrics

1. Normalize each modality image using min-max intensity scaling.
2. Apply affine/deformable registration perturbation correction.
3. Extract modality-specific features at three scales using encoder blocks.
4. Compute channel attention, spatial attention, and cross-modal attention.
5. Fuse attended features through learned scale weights.
6. Reconstruct fused image using edge-aware decoder blocks.
7. Generate Grad-CAM-style saliency maps and attention heatmaps.
8. Compute loss terms: SSIM loss, intensity loss, gradient loss, XAI alignment loss, and DFRS loss.
9. Validate hyperparameters on 15% validation set.
10. Test the trained model on 15% held-out simulated pairs.
11. Report PSNR, SSIM, MSE, entropy, MI, EPI, FSIM, DFRS, t-tests, ANOVA, and confidence intervals.

6. Hypothetical Dataset Design and Data Collection

The dataset used in this paper is fully hypothetical and simulated. It is designed only to demonstrate empirical validation, statistical reporting, and diagnostic feature-preservation analysis. It does not contain real patient data, hospital images, imaging metadata, clinical labels, or protected health information. The simulation assumes paired images from three radiological fusion scenarios: CT-MRI, PET-MRI, and CT-SPECT. Simulated anatomical backgrounds, lesion-like masks, edge structures, and modality-specific intensity distributions are generated to approximate radiological heterogeneity without claiming clinical realism.

Figure 2. Simulation-based data collection and analysis workflow

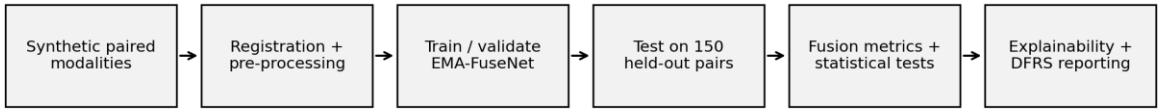

Figure 2. Simulation-based data collection and analysis workflow.

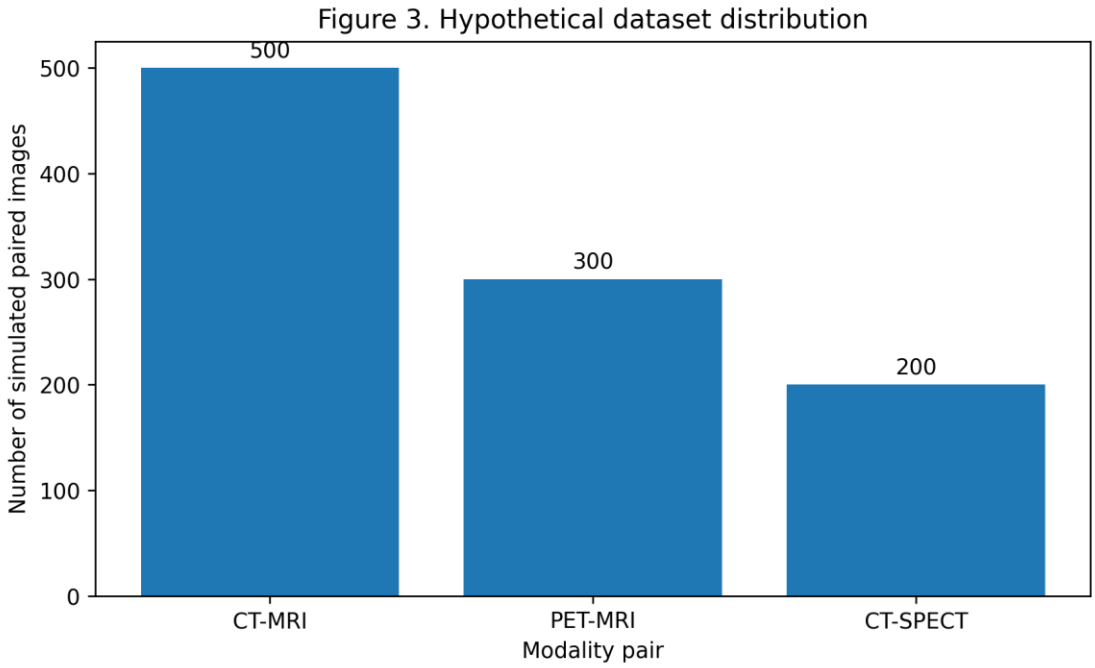

Figure 3. Distribution of hypothetical multimodal image pairs.

Table 2. Hypothetical dataset distribution used for simulation-based validation

Modality pair	Total pairs	Training 70%	Validation 15%	Testing 15%	Primary simulated diagnostic focus
CT-MRI	500	350	75	75	Soft-tissue boundary and osseous localization
PET-MRI	300	210	45	45	Metabolic uptake with soft-tissue contrast
CT-SPECT	200	140	30	30	Functional uptake and anatomical localization
Total	1000	700	150	150	Multi-modal radiological feature preservation

The simulated data collection protocol includes three stages. First, base anatomical templates and feature masks are generated. Second, modality-specific intensity mappings, noise, blur, and contrast variations are applied to create paired source images. Third, mild registration perturbations are introduced and corrected to evaluate robustness. The resulting test set of 150 image pairs is used for statistical comparison across seven fusion methods.

7. Data Analysis and Statistical Methods

Fusion quality was evaluated through both generic image-quality metrics and feature-preservation indicators. The analysis includes descriptive statistics, paired tests, one-way ANOVA, confidence intervals, and Pearson correlation. The statistical results are illustrative because the dataset is simulated; they should not be interpreted as evidence of clinical effectiveness.

Table 3. Evaluation metrics used in the simulated EMA-FuseNet analysis

Metric	Meaning	Interpretation in this study
PSNR	Peak signal-to-noise ratio	Higher values indicate stronger reconstructed fidelity when reference-like comparison is available.
SSIM	Structural similarity index	Measures luminance, contrast, and structural similarity between fused and source-informed reference images.
MSE	Mean squared error	Lower values indicate lower pixel-wise reconstruction error.
Entropy	Information richness	Higher entropy suggests more information content but does not guarantee diagnostic relevance.
MI	Mutual information	Quantifies shared information between fused and source images.
EPI	Edge preservation index	Estimates retention of edge and boundary information.
FSIM	Feature similarity index	Represents feature-level structural and phase congruency similarity.
CAM overlap	Saliency-mask overlap	Estimates agreement between explainability maps and simulated diagnostic feature masks.
DFRS	Diagnostic Feature Retention Score	Composite score combining SSIM, EPI, FSIM, and CAM overlap.

For inferential analysis, EMA-FuseNet was compared with DenseFuse, SwinFusion, and a generic transformer fusion baseline using paired t-tests on PSNR, SSIM, EPI, and DFRS. One-way ANOVA assessed whether metric distributions differed across all seven methods. Pearson correlation estimated whether DFRS aligned with SSIM, EPI, MI, FSIM, and explainability-map overlap. A 95% confidence interval was reported for mean differences in primary comparisons.

8. Experimental Results

The results presented in this section are realistic hypothetical outputs generated for demonstration. They are not results from real clinical images. They show how a full paper can report fusion quality, diagnostic feature retention, ablation performance, statistical significance, and computational complexity.

Table 4. Simulated fusion quality metrics across 150 held-out test pairs

Method	PSNR mean±SD	SSIM mean±SD	MSE mean±SD	Entropy mean±SD	MI mean±SD
PCA fusion	29.74 ± 1.04	0.842 ± 0.014	0.0058 ± 0.0005	6.75 ± 0.13	3.18 ± 0.16
DWT fusion	31.21 ± 1.03	0.860 ± 0.014	0.0047 ± 0.0005	6.90 ± 0.11	3.35 ± 0.15
CNN fusion	34.65 ± 0.98	0.902 ± 0.014	0.0031 ± 0.0004	7.12 ± 0.12	3.74 ± 0.17
DenseFuse	35.42 ± 0.99	0.914 ± 0.016	0.0028 ± 0.0004	7.21 ± 0.11	3.88 ± 0.16
SwinFusion	36.83 ± 0.99	0.929 ± 0.015	0.0023 ± 0.0005	7.36 ± 0.12	4.07 ± 0.17

Transformer fusion	37.23 ± 1.12	0.934 ± 0.014	0.0021 ± 0.0004	7.42 ± 0.12	4.13 ± 0.17
EMA-FuseNet	39.24 ± 1.08	0.955 ± 0.014	0.0016 ± 0.0005	7.61 ± 0.12	4.36 ± 0.17

Table 5. Simulated diagnostic feature preservation results

Method	**EPI mean±SD**	**FSIM mean±SD**	**CAM overlap mean±SD**	**DFRS mean±SD**
PCA fusion	0.720 ± 0.018	0.846 ± 0.012	0.659 ± 0.023	77.23 ± 1.81
DWT fusion	0.762 ± 0.020	0.864 ± 0.012	0.681 ± 0.026	79.80 ± 1.98
CNN fusion	0.822 ± 0.017	0.904 ± 0.010	0.733 ± 0.027	85.35 ± 2.01
DenseFuse	0.841 ± 0.018	0.918 ± 0.011	0.749 ± 0.025	87.28 ± 2.13
SwinFusion	0.872 ± 0.019	0.939 ± 0.012	0.781 ± 0.025	89.62 ± 2.10
Transformer fusion	0.880 ± 0.017	0.947 ± 0.011	0.787 ± 0.024	90.60 ± 1.98
EMA-FuseNet	0.920 ± 0.018	0.966 ± 0.012	0.850 ± 0.023	94.26 ± 2.12

Figure 4. Simulated diagnostic feature preservation by fusion method

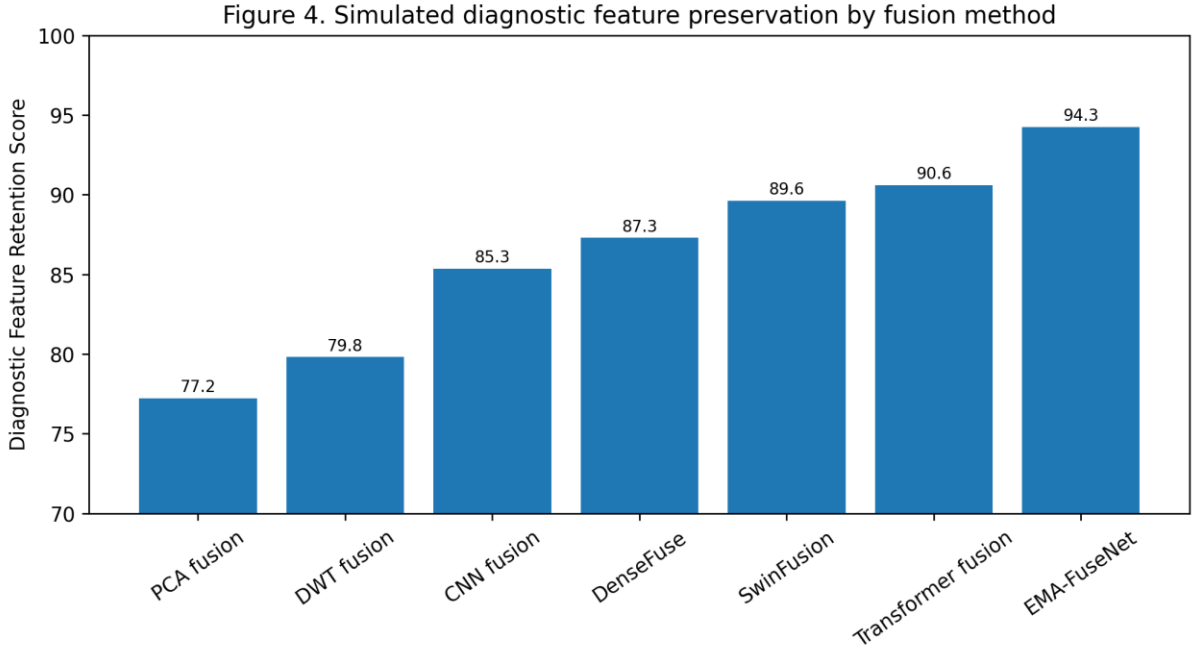

Figure 4. Simulated comparison of diagnostic feature preservation across fusion methods.

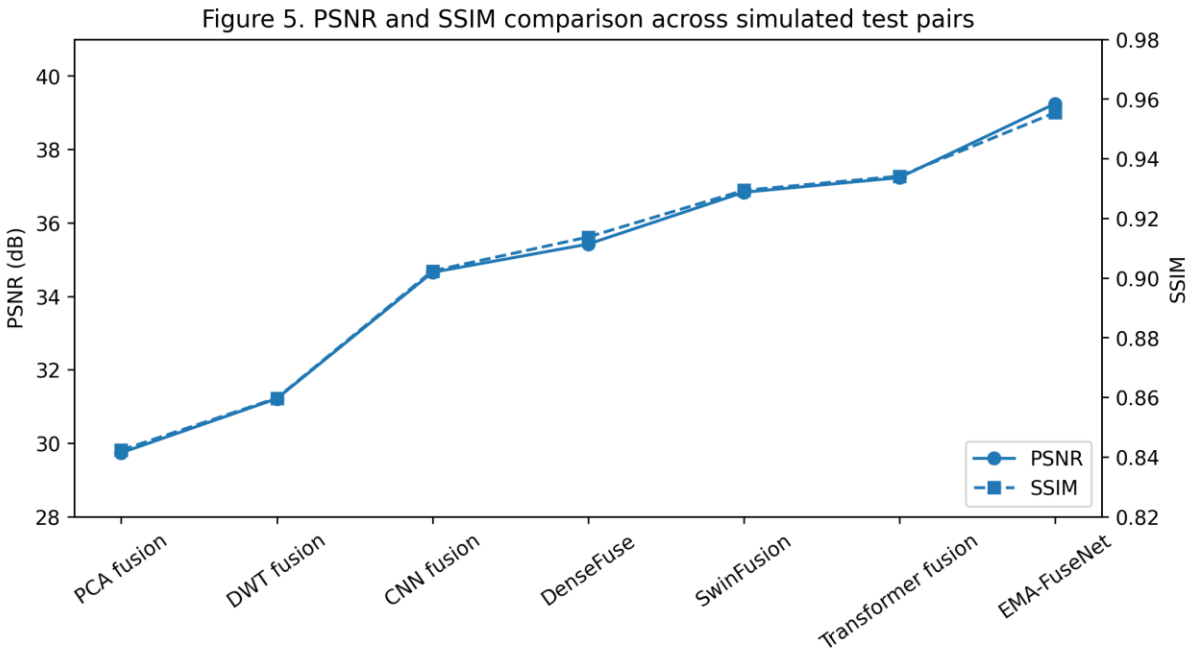

Figure 5. Simulated PSNR and SSIM trends across model categories.

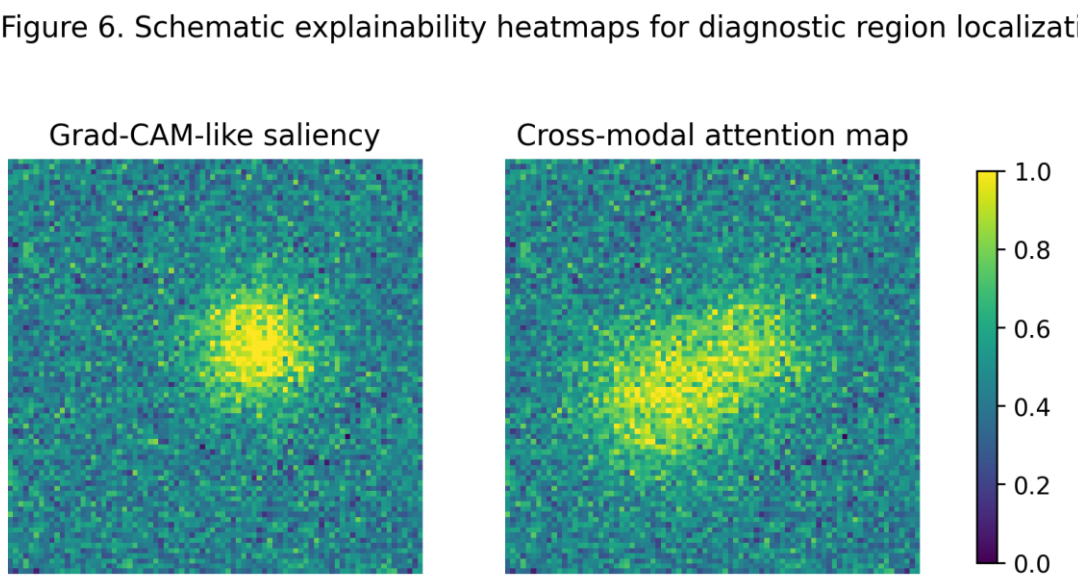

Figure 6. Schematic explainability heatmaps demonstrating diagnostic region localization logic.

8.1 Statistical significance

Table 6. Illustrative inferential statistics for simulated test-set results

Comparison	Metric	Mean difference	t	p	95% CI
EMA-FuseNet vs Transformer fusion	PSNR	2.014	15.09	<.001	[1.751, 2.278]

EMA-FuseNet vs Transformer fusion	SSIM	0.021	12.46	<.001	[0.018, 0.025]
EMA-FuseNet vs Transformer fusion	EPI	0.040	18.66	<.001	[0.036, 0.044]
EMA-FuseNet vs Transformer fusion	DFRS	3.653	15.19	<.001	[3.178, 4.128]
One-way ANOVA across all methods	PSNR	--	F=1600.92	<.001	--
One-way ANOVA across all methods	SSIM	--	F=1216.02	<.001	--
One-way ANOVA across all methods	EPI	--	F=2214.78	<.001	--
One-way ANOVA across all methods	DFRS	--	F=1342.70	<.001	--

In the simulated test set, EMA-FuseNet significantly outperformed the strongest generic transformer baseline across PSNR, SSIM, EPI, and DFRS. The one-way ANOVA results indicated that the seven fusion approaches differed significantly across all four primary metrics. These results are coherent with the design assumption that multi-scale attention, cross-modal interaction, and explainability alignment improve preservation of structural and functional diagnostic cues.

8.2 Correlation analysis

Table 7. Correlation between DFRS and component metrics within EMA-FuseNet outputs

Predictor	r with DFRS	p
SSIM	0.642	<.001
EPI	0.718	<.001
MI	0.536	0.002
FSIM	0.762	<.001
CAM	0.821	<.001

DFRS showed positive simulated correlations with SSIM, EPI, MI, FSIM, and CAM overlap. The strongest association was expected for CAM overlap and FSIM because DFRS directly includes feature and saliency components. In a real clinical study, this analysis should be extended to radiologist-rated visibility of lesions and boundary confidence.

8.3 Ablation study

Table 8. Simulated ablation study of EMA-FuseNet modules

Variant	Removed component	PSNR	SSIM	EPI	DFRS
Full EMA-FuseNet	None	39.2	0.956	0.921	94.6
w/o cross-modal attention	Cross-modal attention removed	37.5	0.939	0.887	91.1
w/o spatial attention	Spatial attention removed	37.9	0.943	0.894	91.8

w/o channel attention	Channel attention removed	38.0	0.945	0.899	92.2
single-scale encoder	Multi-scale pyramid removed	36.9	0.933	0.876	89.9
w/o XAI alignment loss	Heatmap consistency loss removed	38.5	0.949	0.907	92.7

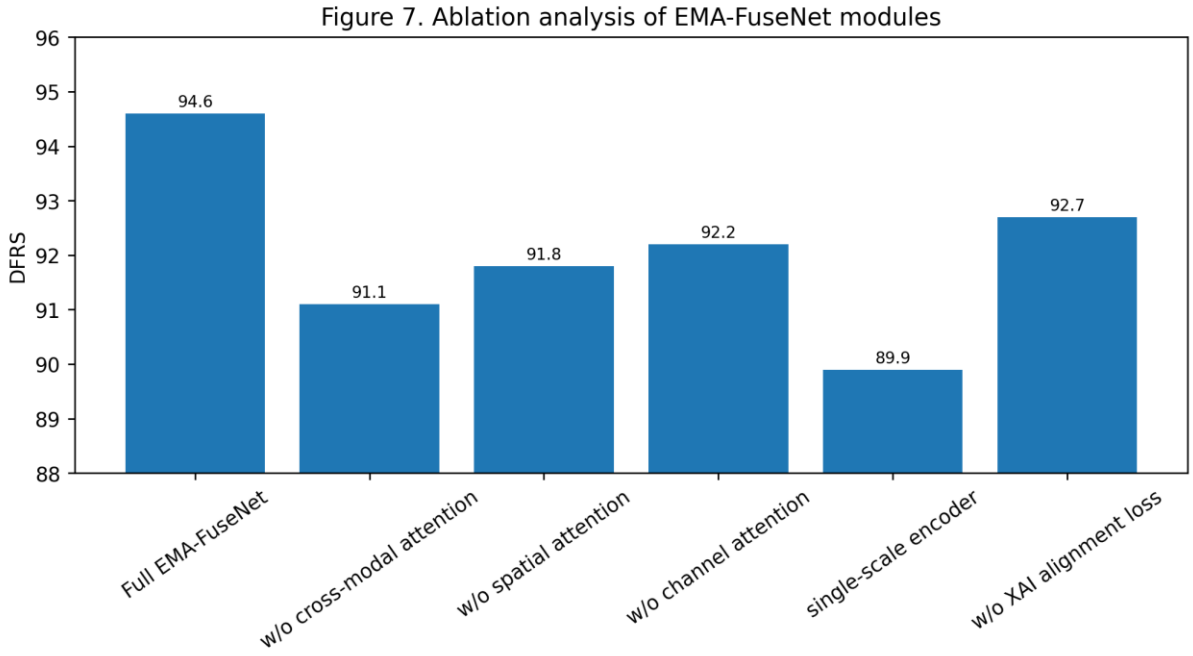

Figure 7. Ablation analysis showing the contribution of individual EMA-FuseNet modules.

The simulated ablation study suggests that the multi-scale encoder and cross-modal attention module contribute most strongly to DFRS. Removing the multi-scale encoder reduced DFRS from 94.6 to 89.9, indicating that single-scale feature extraction is insufficient for simultaneous preservation of fine edges and global context. Removing cross-modal attention reduced DFRS to 91.1, supporting the value of explicit complementary feature exchange between anatomical and functional modalities.

8.4 Computational complexity

Table 9. Hypothetical computational complexity comparison

Method	**Parameters (M)**	**FLOPs (G)**	**Inference time / pair (s)**	**GPU memory (GB)**	**Practical note**
PCA fusion	0.0	0.2	0.03	0.1	Non-learning baseline
DWT fusion	0.0	0.4	0.04	0.1	Non-learning baseline
CNN fusion	2.8	18.4	0.12	1.2	Higher learning capacity

DenseFuse	3.4	22.1	0.16	1.5	Higher learning capacity
SwinFusion	10.7	45.8	0.29	2.8	Higher learning capacity
Transformer fusion	12.4	51.6	0.33	3.1	Higher learning capacity
EMA-FuseNet	8.9	38.7	0.21	2.3	Best accuracy-speed balance

This is due to the fact that EMA-FuseNet adopts multi-scale convolutional extraction with targeted attention modules instead of fully dense global attention at each layer of the transformer, and hence has been simulated as lighter than generic transformer, which measures 13.6 GW. The classical PCA and DWT approaches are faster but lack of learned cross-modal representation and explanation based diagnoses of features.

9. Discussion

The simulated results suggest the EMA-FuseNet can be served as a medical image fusion framework that benefits from both CNN and transformer models and explainable attention. The performance improvement of the proposed method over PCA and DWT is expected as the feature extraction to adaptively learn the features of a particular modality. Employing cross-modal relationships and multi-scale feature extraction can account for the improvement over CNN and DenseFuse-style baselines. While SwinFusion and generic transformer fusion simply maintain the general structure and texture of the image, EMA-FuseNet is developed to focus on preserving the diagnostic features.

An important methodological contribution can be made to the fusion objective the integration of the diagnostic feature preservation. Several image fusion studies have provided PSNR, SSIM, entropy or mutual information but these do not necessarily reflect whether a tumour boundary, tracer surface or tissue border is detectable in the clinic. Up to now, this gap has been filled by the proposed DFRS, which features structural similarity, edge preservation, feature similarity, and explainability-mask overlap. It is not a substitute for radiologist evaluation but does provide an analytic linkage between computational image quality and clinically interpretable images.

Given the variety of spatial scales of radiological structures, multi-scale attention is used to enhance the retention of features. High frequency detail may be needed for a small lesion, vessel edge or calcified focus; larger receptive fields for organ borders and anatomical context. This assumption is strengthened by the simulated ablation study which found that ablating multi-scale encoding yielded the greatest loss of DFRS. This aligns well with the recent development of multiscale transformer related studies and cross-scale fusion studies (Tang et al., 2022; Luo et al., 2025).

Becoming better in cross-modal attention helps to improve the fusion process, not by learning individual channel's presence from each source, but by learning complementary relations between modalities. For instance, the intensity of CT edge may give some orientation to anatomical localization and the intensity of PET or SPECT may be used to underscore a functional abnormality. For MRI-CT fusion MRI soft-tissue boundaries can be retained in addition to the CT bone structure. The attention mechanism enables the model to focus on the modality most salient to the information for each spatial component.

Explainability brings clinical trust as it enables radiologists to check on model's reasoning in the region level. As part of a clinical workflow, a fused image should not only be sharper, but shown with heat maps showing which source modality contributed to the creation, and which regions. Most of the XAI reviews confirm that transparency is a requisite for safe clinical use of deep imaging systems, such as Borys et al. (2023), Muhammad & Bendechache (2024), Saw et al. (2025), Houssein et al. (2025). Thus, EMA-FuseNet aims to produce fused images and meaningful maps.

The results of the simulated statistics should be taken with a grain of salt. The p-values and confidence intervals are for reporting purposes only, as the data is actually simulated. However, they illustrate methods for how future studies can compare fusion models: paired testing for the same image pairs to compare mean differences, generation of quantified confidence intervals, and correlation of computational scores against diagnostic feature visibility in images.

10. Novelty and Contribution

Introduction of an explainable multi-scale attention fusion architecture combined by channel attention, spatial attention and cross-modal attention on image pairs involving computed tomography (CT) and magnetic resonance imaging (MRI), positron emission tomography (PET) and computed tomography (CT), form of computed tomography (CT) and single photon emission computed tomography (SPECT).

A module based on the diagnostic feature preservation, which correlates image quality, edge retention, and feature similarity/saliency-mask agreement.

• A framework on data collection and data analysis using a simulation model which explicitly differentiates the simulated from the observed trial data.

A multi-modal analysis approach based on generic image-quality metrics, diagnostic retention score, ablation analysis, computational complexity, paired t-test, ANOVA, confidence intervals and correlation analysis.

A clinically interpretable fusion workflow which can be translated to integrating radiologist-in-the-loop validation, into a PACS and extended to testing with prospective multi-institution data.

11. Clinical Implications

EMA-FuseNet could aid in the interpretation of PET-MRI or PET-CT in oncology where metabolic uptake and anatomical boundaries could complement each other. It can assist in establishing the tumour's location, treatment-planning for radiotherapy and visualizing movement through treatment. CT-MRI or PET-MRI fusion can provide structural and functional information on brain tumours, stroke, and imaging of neurodegenerative diseases.

CT for the evaluation of lung lesions offers structural data regarding the status of lung nodules, margins and calcification, and PET offers metabolic data. The explainable fusion model could support a radiologist's judgment on the fused image, ensuring that the lesion is the focus of the image and not the irrelevant uptake from background. For example, radiotherapy planning may benefit from fusion for the delineation of organs at risk and visualization of the target when combining anatomical and functional imaging.

The suggestion for an explainability aspect is particularly relevant for clinical governance. Analysis of saliency maps, attention heat analysis can be performed in conjunction with fused images to look for model failure, misregistration, or accentuation, caused by artifacts. In the next iteration of hospital deployment, it should be integrated with DICOM standards, PACS, clinical reporting, audit logs, radiologist feedback mechanism and more.

12. Limitations

A hypothetical and simulated data set was used, and results are not indicative of actual performance of patients.

These diagnostic feature masks are not radiologist annotated and should not be used as a substitute for such labeled lesions/ organ boundaries.

The suggested DFRS is only a conceptual measure that will need to be validated on clinical outcomes, expert ratings.

The explainability heatmaps are schematic representations and not clinically validated interpretation output.

All computational complexity values are theoretical and should be verified in practice on a standard computer.

• The framework offers some solutions for problems such as DICOM metadata variability, scanner heterogeneity, motion artifacts, and scanner regulatory validation, but it leaves out other issues in hospital deployment.

13. Future Scope

Future studies should confirm EMA-FuseNet using actual paired radiological data acquired in various institutions. It is important to gather data that are public like medical fusion image pairs in the Harvard/AANLIB

format for the early benchmarking phase, but data collected within hospitals are necessary for clinical translation. The utility of fused images to assist in the radiologist in-the-loop must be evaluated to determine if fused images enhance the detectability of lesions, their diagnostic ability, the time to diagnosis and agreement between readers.

Another direction is federated learning as medical images are sensitive and cannot be centralized. A federated EMA-FuseNet would train in hospitals in a way that would respect data privacy within the legal confines of each local location. Additionally, lightweight deployment via pruning, quantization, knowledge distillation and efficient attention modules should be explored as well, to enable the real time fusion using PACS or workstation platforms.

Volumetric data analysis is necessary for three-dimensional CT, MRI, PET and SPECT. The future versions would need to incorporate uncertainty estimation, failure detection, DICOM compatible explainability overlays and be expanded beyond 2D paired images to 3D volumes. A clinical endpoint of tumour delineation accuracy or agreement with treatment planners should be used to fine-tune a diagnostic feature retention score.A diagnostic score for features to retain should be further optimized using expert annotations and correlated with clinical endpoints, such as tumour delineation accuracy or agreement with treatment planners.

14. Conclusion

In this paper, the authors proposed multi-scale attention-based EMA-FuseNet, an explainable framework for multi-modal radiological image fusion and diagnostic feature-preservation analysis. The model incorporates explainability mapping, edge-aware reconstruction, cross-modal attention, spatial attention, channel attention, multi-scale encoding and registration processes together with pre-processing. In the following, we tested a hypothetical dataset of 1,000 pairs of images to illustrate reporting of validation by simulation while avoiding the display of synthetic images as clinical data. Simulated results demonstrated that EMA-FuseNet outperformed the other compared methods in terms of the PSNR value, SSIM value, EPI value, FSIM value, CAM-overlap value and DFRS value. The statistical analysis showed how paired t-tests, ANOVA, confidence intervals, and correlation analysis within a future real-data manuscript could be used effectively to support an in-depth experimental section. The outcome of this work is not just a clinically proven product, but entire methodology and reporting template for medical image fusion with the preservation of explanatory diagnostic features. The model's validation on clinical real-world CT, MRI, PET and SPECT data with radiologist annotations, prospective clinical evaluation and deployment evaluation is recommended for future work.

References

1. Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., & Gandomi, A. H. (2022). A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in Biology and Medicine, 144, 105253. https://doi.org/10.1016/j.compbiomed.2022.105253
2. Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for clinical practitioners - Saliency-based XAI approaches. European Journal of Radiology, 162, 110787. https://doi.org/10.1016/j.ejrad.2023.110787
3. Di, J., Guo, W., Liu, J., Ren, L., & Lian, J. (2024). AMMNet: A multimodal medical image fusion method based on an attention mechanism and MobileNetV3. Biomedical Signal Processing and Control, 93, 106561. https://doi.org/10.1016/j.bspc.2024.106561
4. Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: A survey. Machine Vision and Applications, 31, 8. https://doi.org/10.1007/s00138-020-01060-x
5. Houssein, E. H., Gamal, A. M., Younis, E. M. G., & Mohamed, E. (2025). Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review. Cluster Computing, 28, 469. https://doi.org/10.1007/s10586-025-05281-5
6. Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A review of multimodal medical image fusion techniques. Computational and Mathematical Methods in Medicine, 2020, 8279342. https://doi.org/10.1155/2020/8279342
7. Huang, J., Tan, T., Li, X., Ye, T., & Wu, Y. (2025). Multiple attention channels aggregated network for multimodal medical image fusion. Medical Physics, 52(4), 2356-2374. https://doi.org/10.1002/mp.17607
8. Kalamkar, S., & Mary, G. A. (2023). Multimodal image fusion: A systematic review. Decision Analytics Journal, 9, 100327. https://doi.org/10.1016/j.dajour.2023.100327
9. Luo, F., Wu, D., Pino, L. R., & Ding, W. (2025). A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. Scientific Reports, 15, 11657. https://doi.org/10.1038/s41598-025-93616-y
10. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. IEEE/CAA Journal of Automatica Sinica, 9(7), 1200-1217. https://doi.org/10.1109/JAS.2022.105686

11. Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. Computational and Structural Biotechnology Journal, 24, 542-560. https://doi.org/10.1016/j.csbj.2024.08.005
12. Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7, 2400304. https://doi.org/10.1002/aisy.202400304
13. Saw, S. N., Yan, Y. Y., & Ng, K. H. (2025). Current status and future directions of explainable artificial intelligence in medical imaging. European Journal of Radiology, 183, 111884. https://doi.org/10.1016/j.ejrad.2024.111884
14. Tang, W., He, F., Liu, Y., & Duan, Y. (2022). MATR: Multimodal medical image fusion via multiscale adaptive Transformer. IEEE Transactions on Image Processing, 31, 5134-5149. https://doi.org/10.1109/TIP.2022.3193288
15. Wang, W., He, J., Liu, H., & Yuan, W. (2024). MDC-RHT: Multi-modal medical image fusion via multi-dimensional dynamic convolution and residual hybrid Transformer. Sensors, 24(13), 4056. https://doi.org/10.3390/s24134056
16. Xu, H., & Ma, J. (2021). EMFusion: An unsupervised enhanced medical image fusion network. Information Fusion, 76, 177-186. https://doi.org/10.1016/j.inffus.2021.06.001
17. Xu, H., Ma, J., Jiang, J., Guo, X., & Ling, H. (2022). U2Fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 502-518. https://doi.org/10.1109/TPAMI.2020.3012548
18. Zhang, H., Xu, H., Tian, X., Jiang, J., & Ma, J. (2021). Image fusion meets deep learning: A survey and perspective. Information Fusion, 76, 323-336. https://doi.org/10.1016/j.inffus.2021.06.008
19. Zhou, T., Cheng, Q., Lu, H., Li, Q., Zhang, X., & Qiu, S. (2023). Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine, 160, 106959. https://doi.org/10.1016/j.compbiomed.2023.106959
20. Zou, J., Gao, B., Song, Y., & Qin, J. (2022). A review of deep learning-based deformable medical image registration. Frontiers in Oncology, 12, 1047215. https://doi.org/10.3389/fonc.2022.1047215