DEEP LEARNING-DRIVEN MULTI-MODAL MEDICAL IMAGE FUSION USING CNN FOR ENHANCED BRAIN TUMOR DIAGNOSIS

[1]A.Yohan, [2]Mr.P. Surya, [3]M.Venkatesh, [4]K.Nithin Reddy, [5]K.Bala Adharsh,[6]Y.Jagan Mohan Reddy

[2]Associate Professor, ECE Dept, RISE Krishna Sai Prakasam Group of Institutions, Valluru, AP

[1,3,4,5,6]Students, ECE Dept, RISE Krishna Sai Prakasam Group Of Institutions, Valluru, AP

[1]mr.rikayohan@gmail.com,[2]suryame3020@gmail.com,[3]adarshkunchala7@gmail.com,[4]kanulanithin0602@gmail.com, [5] jaganmohanreddyyannam79@gmail.com, [6]vmayaluri9@gmail.com

Submitted: 05-01-2026 *Accepted: 13-02-2026* *Published: 20-02-2026*

ABSTRACT

Imaging methods like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single-Photon Emission Computed Tomography (SPECT) have given medical professionals information about the soft tissues, structural features, and other aspects of the human body. distinct sensors collect distinct image information of the same part, and different imaging techniques maintain different properties. Better contrast, fusion quality, and perceived experience are the goals of the fusion. The outcome of the fusion should satisfy the following requirements: (a) the fused image should fully preserve the information of the source images; (b) the fused image should not generate any artificial information, such as artifacts; and (c) poor states, such as noise and misregistration, should be avoided. The suggested network is one of three CNN models used to compare patch similarity. Because its two weight branches are the same, the source image's feature extraction and activity level measuring methodologies are also same. This has some advantages over pseudo and 2-channel models, and the siamese model is also used in fusion applications due to its ease of training. After getting the weight map, the Gaussian pyramid decomposition is used, followed by the pyramid transform for multiscale decomposition, bringing the fusion process closer to human visual perception. Additionally, the decomposed coefficients are adaptively adjusted using the localized similarity-based fusion technique. The algorithm combines the popular pyramid-based and similarity-based fusion algorithms with the CNN model to provide a better fusion approach

I INTRODUCTION

Medical imaging is essential for clinical diagnosis, treatment planning, and surgical navigation, thanks to rapid improvements in sensor and computer technology. Because different imaging modalities record different sorts of anatomical and functional information, numerous modalities such as X-ray, CT, MR, PET, and SPECT are frequently used in clinical settings. CT pictures are especially good at viewing solid structures like bones and implants, whereas MR imaging reveal high-resolution soft-tissue features. PET and SPECT, on the other hand, are more focused on functional information like blood flow and metabolic changes, but with poorer spatial resolution. Multi-modal medical image fusion combines complimentary information from various modalities to create a single composite image, allowing for clearer visualization and assisting clinicians in making more accurate and efficient clinical judgments.

II LITERATURE SURVEY

Significant studies have been performed in the last few years to understand the application of deep learning algorithms in medical image fusion. These initiatives have been led by Convolutional Neural Networks (CNNs) because of their ability to learn hierarchical spatial features from pictures [1]. As it has previously been established, models whose architecture is based on CNNs have exceptional effectiveness when solving medical picture segmentation, classification, and fusion problems. Their ability to glean specific, high-resolution specifics, for instance, textures and edges,[2] has made them a go-to choice for medical image analysis. However, despite their excellent performances in several tasks, CNNs suffer from inherently limited receptive fields [3]. This makes it

difficult for CNN-based models to minimize the dependencies of far pixels, which is usually appropriate for analysing complex medical pictures such as mammography for tumour detection [4]. In recent advancements, Vision Transformers (ViTs) have surfaced as a new way of capturing image information. Unlike CNNs, where features are captured with convolutional filters that analyse the local region of the image, ViTs use self-attention procedures that allow the model to view relations across the entire image [5]. Overall, ViTs are particularly good at capturing global contextual information and understanding the whole picture of an image. Generally, Vision [6]. Transformers (ViTs) have demonstrated high performance in various computer vision tasks such as object detection, image classification, and synthesis results. However, while ViTs excel in capturing global dependencies, they perform poorly in capturing local patterns – critical for high-precision tasks like delineating tumour margins in medical images. Hybrid models incorporating the strengths and limitations of CNNs and ViTs have recently been proposed. A compound use of CNNs and ViTs may offer a superior understanding of a medical picture since CNNs excel in local feature extraction [7].

III EXISTING METHOD

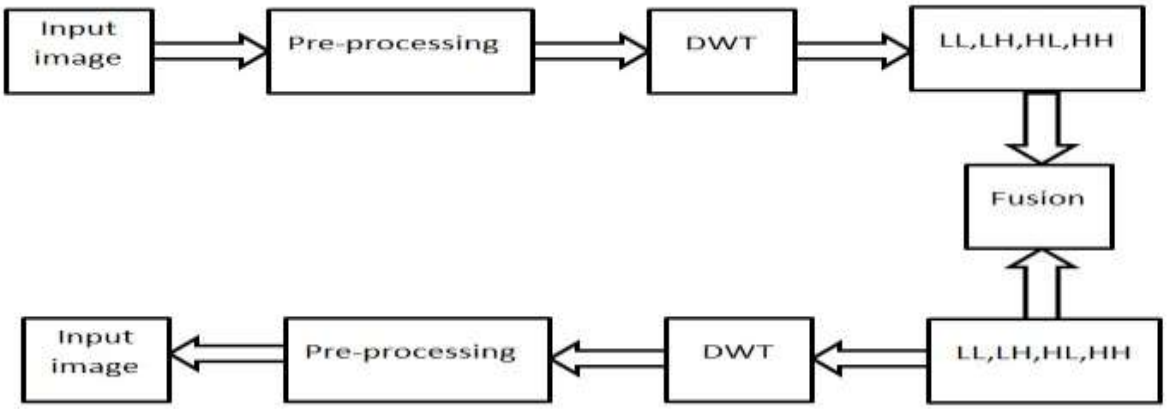

Fig 1: Block diagram of Existing Model

Input Image:

The input image is the original image used in the fusion process, which is typically obtained from multiple sensors or imaging modalities. Each image provides distinct and complimentary information that may not be fully represented in the other images. These photographs are the raw data for the fusion system. The quality and qualities of the input photos have a direct impact on the success of the fusion process.

Pre-processing:

Pre-processing is used to improve the quality of the input photos prior to further analysis. This stage may include noise reduction, contrast improvement, image normalization, and scaling. It also guarantees that all supplied photographs follow the same format and resolution. Proper pre-processing aids in the extraction of significant information while also improving fusion accuracy.

DWT (Discrete Wavelet Transform):

The Discrete Wavelet Transform divides the pre-processed image into several frequency components. It allows the image to be evaluated at various resolutions. DWT effectively separates critical image information, such as edges and textures, from smooth regions. Its multi-resolution capability makes it ideal for picture fusion applications.

LL, LH, HL, HH Sub-bands:

Following DWT, the image is separated into four sub-bands: LL, LH, HL, and HH. The LL subband comprises low-frequency information that describes the image's general structure and lighting. The LH, HL, and HH subbands contain high-frequency details like vertical, horizontal, and diagonal edges. These sub-bands aid in selectively combining relevant elements from each image.

Fusion:

Fusion is the technique of joining corresponding sub-bands from numerous images to create a single set of fused coefficients. Different fusion rules are used to extract the most useful information from each image. This stage improves key characteristics while minimizing redundancy. The fused sub-bands are then used to create a final enhanced image via inverse DWT.

IV PROPOSED METHOD

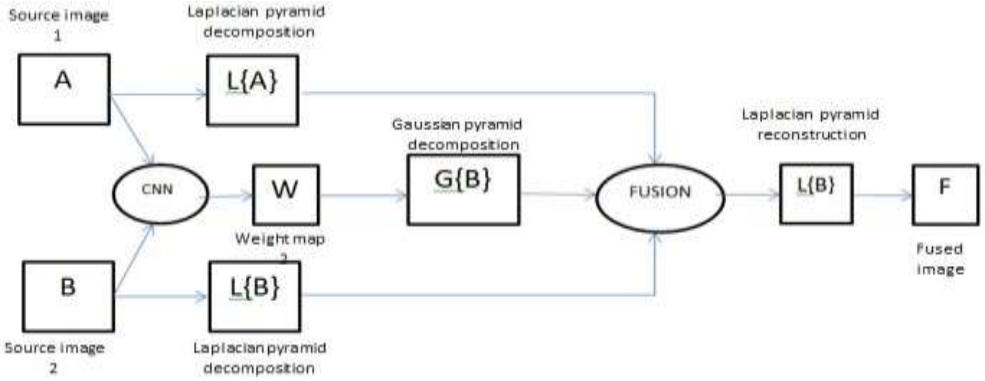

Fig 2: Block diagram of Proposed Model

Source Image A and Source Image B:

Source images A and B are the original input images obtained using various sensors, perspectives, or imaging modalities. Each image contains complimentary information, such as differences in texture, focus, lighting, and spectral content. These photographs are the key data sources for the fusion framework. Processing them together helps the system to capitalize on their individual capabilities. The ultimate goal is to combine meaningful information from both photos into a single representation. The quality and properties of these source photos have a considerable impact on the success of the fusion process. Both photos are processed independently at the beginning stages to preserve their unique traits.

Laplacian Pyramid Decomposition (L{A} and L{B}):

Laplacian pyramid decomposition is used on both source images to represent them at various resolution levels. This approach divides the image into band-pass components, which highlight edges, textures, and small details. High-frequency layers capture dramatic transitions and structural details, whereas lower levels depict softer changes. Such breakdown aids in identifying critical traits that must be conserved during fusion. It also lowers redundancy at different sizes and improves multi-resolution analysis. Laplacian pyramids are commonly utilized because they efficiently preserve spatial features. This makes them ideal for image fusion tasks.

CNN (Convolutional Neural Network):

The convolutional neural network uses the original images to learn deep and discriminative features. It captures relevant patterns such as edges, textures, and prominent regions without requiring operator interaction. The CNN adapts to a variety of image features and imaging situations. Its output helps discover regions that are more informational or visually significant. These learned characteristics improve the correctness of the fusion judgment. Using a CNN makes the fusion structure more resilient and flexible. This method often generates better visual and perceptual fusion results.

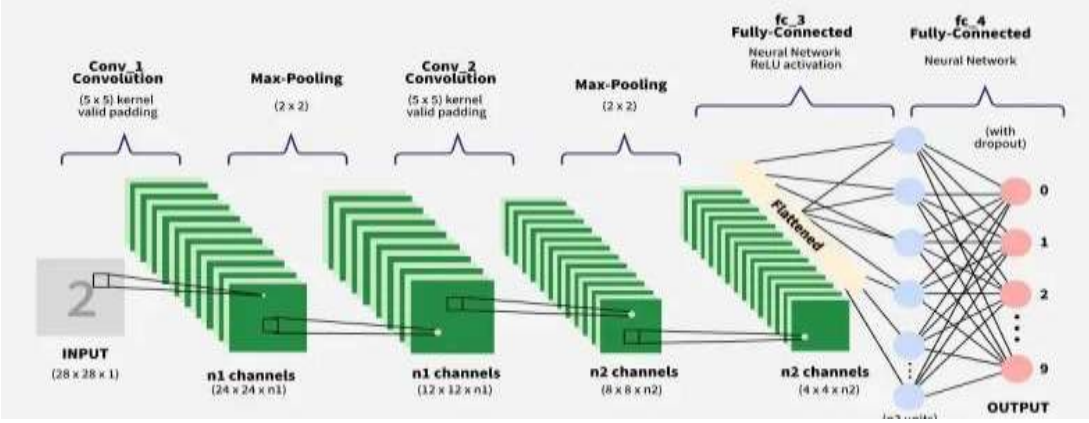

Fig 3: CNN output of various convolution stages

Weight Map (W):

The weight map represents the relative importance of each pixel or region in the source images and it is generated using the feature information obtained from the CNN. Regions with higher weights contribute more strongly to the fusion process allowing the system to preserve dominant and informative features from each image. The weight map varies spatially allowing adaptive fusion across the image and it helps to reduce artifacts and preserve fine details.

Gaussian Pyramid Decomposition (G{B}):

By repeatedly smoothing and down sampling the image, the Gaussian pyramid decomposition produces a multi-scale representation. It primarily gathers low-frequency and structural information from the image. This decomposition compliments the Laplacian pyramid, which emphasizes detail information. Gaussian pyramids are frequently used to smoothly blend data across scales. They aid in uniformity during fusion and repair. This strategy lowers noise and prevents abrupt transitions. Gaussian decomposition is required for consistent and visually appealing **fusion outcomes. Fusion:**

The fusion stage integrates data from the Laplacian and Gaussian pyramids of both pictures. Using the weight map, the respective pyramid levels are combined adaptively. Important characteristics of each source image are selectively retained. Fusion on multiple sizes ensures that both global structures and tiny details are preserved. This approach minimizes redundancy while increasing relevant information. It improves the overall visual quality of the image. This stage produces a collection of fused pyramid coefficients.

Laplacian Pyramid Reconstruction:

Laplacian pyramid reconstruction is the inverse of decomposition. The fused pyramid coefficients are merged to reconstruct the image. This includes up sampling and filtering at each pyramid level. Reconstruction ensures that data from all scales are appropriately merged. Proper reconstruction prevents visual artifacts and distortions. It ensures spatial uniformity in the final image. This step is necessary to provide a high-quality fused **outcome. Fused Image (F):**

The fused image is the final output obtained after reconstruction. It contains enhanced information from both source images. Important features such as edges, textures, and structural details are preserved. The fused image provides improved clarity and interpretability. It is more informative than either source image alone. Such images are useful in applications like medical imaging, remote sensing, and surveillance. The quality of the fused image reflects the effectiveness of the fusion method.

V ANALYSIS OF PERFORMANCE AND RESULTS

The CT image primarily displays solid structures such as bone with strong contrast and crisp edges, whereas the MR-Gad image emphasizes soft tissue and vascular features due to contrast enhancement, offering complementing but incomplete information. The fused output retains the structural clarity of CT as well as the soft-tissue information of MR, resulting in a more informative and balanced image. Bone features stay sharp, interior brain tissues appear clearer, and overall contrast improves when compared to either source alone. This demonstrates that

the fusion procedure successfully preserves key elements from both modalities. Such a composite image can help with clinical interpretation and potentially enhance diagnosis accuracy.

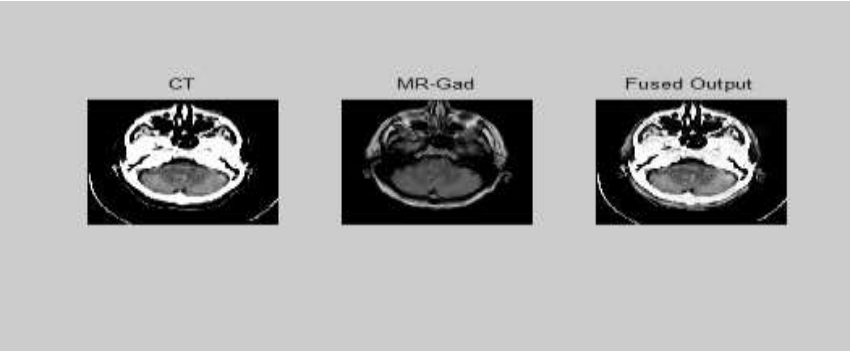

Fig 4: Results of fused output

CONFUSION MATRIX:

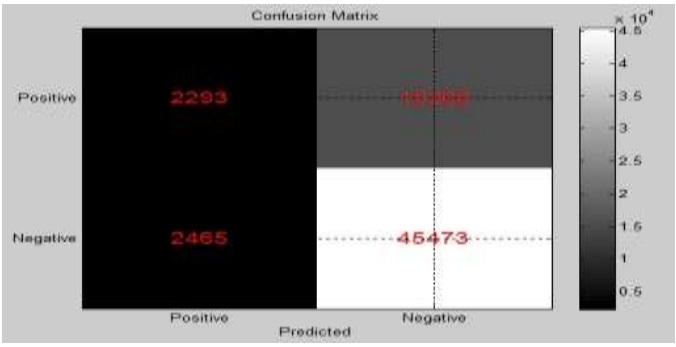

Fig 5: Performance analysis of Confusion matrix output

The confusion matrix illustrates the model's classification performance for two classes, revealing a significant bias toward predicting the negative class. There are 2293 true positives, which means that some positive cases are accurately detected; however, a huge number of positives (15305) are overlooked, showing low sensitivity or recall for the positive class. The approach also generates 2465 false positives (negatives falsely labeled as positive), whereas true negatives (45473) dominate the results. This shows that the dataset is severely skewed, with far more negative samples than positive ones. Although the overall accuracy appears to be high due to the vast number of genuine negatives, the model's ability to detect positive cases is very weak and requires development.

PERFORMANCE MATRICS COMPARISON:

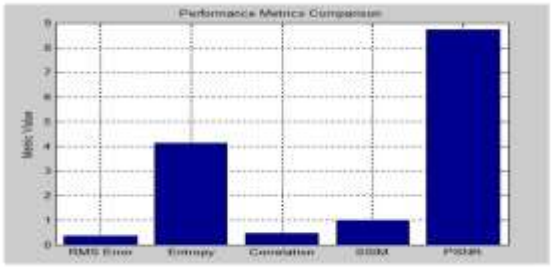

Fig 6: Performance Metrics Comparison

The bar chart contrasts key performance measures used to assess the quality of a processed or fused image. The RMS Error is low, indicating that the total difference between the reference and processed images is minor, and the reconstruction error is minimal. Entropy has a moderate value, indicating that the image includes a sufficient quantity of information and detail without being unduly noisy. The correlation value is in the middle of the range,

indicating that the input and output images are structurally comparable. The SSIM result is near to one, indicating high structural similarity and retention of key picture elements. PSNR is the greatest of the measures, indicating good signal quality and low noise in the output image. A higher PSNR typically indicates greater visual quality and fewer aberrations. Low RMS error and high PSNR suggest that the fusion or processing approach is effective. The SSIM test confirms this, demonstrating that structural features are substantially maintained. The modest entropy indicates a balance between detail improvement and noise control. Overall, the criteria indicate that the generated image has good fidelity, structure, and visual quality when compared to the reference.

VI CONCLUSION

This study offers a successful multi-modal medical image fusion method that leverages the complementary characteristics of CT and MR images. The suggested technique effectively maintains structural details from CT and soft tissue information from MR, resulting in a more informative fused image. Visual results demonstrate better contrast, clarity, and feature representation than individual source photos. Quantitative criteria such as low RMS error and high PSNR indicate strong reconstruction quality and low noise. A high SSIM value suggests good structural similarity and consistent feature preservation. Although the confusion matrix reveals a class imbalance and low sensitivity for positive cases, the whole architecture has great potential. The fusion technique based on CNN and multi-scale pyramids is both robust and flexible. It minimizes artifacts and improves clinically significant details. This improves the diagnostic and interpretability of the merged pictures. Overall, the suggested method represents a viable alternative for improved medical picture analysis and decision assistance.

REFERENCES:

[1] A. James and B. Dasarathy, “Medical image fusion: a survey of the state of the art,” Information Fusion, vol. 19, pp. 4–19, 2014.

[2] L. Yang, B. Guo, and W. Ni, “Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform,” Neurocomputing, vol. 72, pp. 203–211, 2008.

[3] Z. Wang and Y. Ma, “Medical image fusion using m-pcnn,” Information Fusion, vol. 9, pp. 176–185, 2008.

[4] B. Yang and S. Li, “Pixel-level image fusion with simultaneous orthogonal matching pursuit,” Information Fusion, vol. 13, pp. 10–19, 2012.

[5] S. Li, H. Yin, and L. Fang, “Group-sparse representation with dictionary learning for medical image denoising and fusion,” IEEE Transactions on Biomedical Engineering, vol. 59, pp. 3450–3459, 2012.

[6] Z. L. G. Bhatnagar, Q. Wu, “Directive contrast based multimodal medical image fusion in nsct domain,” IEEE Transactions on Multimedia, vol. 15, pp. 1014–1024, 2013.

[7] S. Li, X. Kang, and J. Hu, “Image fusion with guided filtering,” IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2864–2875, 2013.

[8] R. Shen, I. Cheng, and A. Basu, “Cross-scale coefficient selection for volumetric medical image fusion,” IEEE Transactions on Biomedical Engineering, vol. 60, pp. 1069–1079, 2013.

[9] R. Singh and A. Khare, “Fusion of multimodal medical images using daubechies complex wavelet transform c a multiresolution approach,” Information Fusion, vol. 19, pp. 49–60, 2014.

[10] L. Wang, B. Li, and L. Tan, “Multimodal medical volumetric data fusion using 3-d discrete shearlet transform and global-to-local rule,” IEEE Transactions on Biomedical Engineering, vol. 61, pp. 197–206, 2014.