An Ensemble Learning Approach for Multi-Modal Medical Image Fusion using Deep Convolutional Neural Networks

Andino Maseleno[1], Dr. D. Kavitha[2], Koudegai Ashok[3], Dr. Mohammed Saleh Al Ansari[4],
Nimmati Satheesh[5], Dr. R. Vijaya Kumar Reddy[6]
Institut Bakti Nusantara, Lampung, Indonesia[1]
Associate Professor, Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada[2]
Associate Professor, Vignana Bharathi Institute of Technology, Ghatkesar, Hyderabad[3]
Associate Professor, College of Engineering, Department of Chemical Engineering, University of Bahrain, Bahrain[4]
Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul[5]
Associate professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India[6]

Abstract*—Medical image fusion plays a vital role in enhancing the quality and accuracy of diagnostic procedures by integrating complementary information from multiple imaging modalities. In this study, we propose an ensemble learning approach for multi-modal medical image fusion utilizing deep convolutional neural networks (DCNNs) to predict brain tumour. The proposed method aims to exploit the inherent characteristics of different modalities and leverage the power of CNNs for improved fusion results. The Generative Adversarial Network (GAN) strengthens the input images. The ensemble learning framework comprises two main stages. Firstly, a set of DCNN models is trained independently on the respective input modalities, extracting high-level features that capture modality-specific information. Each DCNN model is fine-tuned to optimize its performance for fusion. Secondly, a fusion module is designed to aggregate the individual modality features and generate a fused image. The fusion module employs a weighted averaging technique to assign appropriate weights to the features based on their relevance and significance. The fused image obtained through this process exhibits enhanced spatial details and improved overall quality compared to the individual modalities. On a diversified dataset made up of multi-modal medical images, thorough tests are carried out to assess the efficacy of the suggested approach. The fusion images exhibit improved visual quality, enhanced feature representation, and better preservation of diagnostic information. The BRATS 2018 dataset, which contains Multi-Modal MRI images and patients' healthcare information were used. The proposed method also demonstrates robustness across different medical imaging modalities, highlighting its versatility and potential for widespread adoption in clinical practice.

***Keywords*—Deep convolutional neural networks; image fusion; generative adversarial network; ensemble learning**

I. Introduction

In recent years, the field of medical imaging has witnessed tremendous advancements with the availability of multiple imaging modalities. Each modality provides unique information about anatomical structures, functional processes, or disease characteristics, making it crucial to extract comprehensive insights by combining data from multiple modalities [1]. To effectively utilize the complementary information present in multi-modal medical images, researchers have turned to image fusion techniques. Image fusion aims to integrate data from different modalities into a unified representation, allowing for enhanced visualization, improved diagnostic accuracy, and better decision-making in clinical settings [2]. Several tasks related to computer vision, such as picture categorization, object identification, and categorization, have shown DCNN to be quite effective. In recent years, DCNNs have also gained significant attention in the domain of medical image analysis due to their ability to automatically learn complex features from large-scale data [3].

With the use of convolutional neural networks with deep layers, this work suggests a collaborative learning approach for fusing multimodal medical images. This method aims to overcome the limitations of traditional fusion methods by automatically learning the optimal fusion strategy from the data itself. The ensemble learning framework involves training multiple DCNNs, each specializing in capturing distinct features from different modalities [4]. These networks are designed to learn a shared representation that combines the information from each modality effectively. The ensemble is formed by aggregating the predictions of these networks, yielding a fused image that encapsulates the strengths of each modality [5]. By adopting an ensemble approach, this method leverages the diversity and complementary nature of the individual networks, resulting in a more robust and accurate fusion outcome. Additionally, the ensemble enables us to address uncertainties associated with the fusion process by providing a measure of confidence for the final fused image [6].

We carried out comprehensive experiments on a wide range of multi-modal healthcare imaging datasets in order to assess the performance of the suggested technique. The results demonstrate the superiority of ensemble learning approach over traditional fusion methods and even single DCNN-based fusion techniques. The fused images exhibit improved clarity, enhanced structural details, and better discrimination of abnormal regions, making them highly valuable for clinical decision support and medical research. ensemble learning approach for multi-modal medical image fusion, employing deep convolutional neural networks, offers a promising solution for extracting comprehensive information from multi-modal medical images [7]. By effectively integrating the strengths of different imaging modalities, this method holds the potential to advance the field of medical imaging and facilitate more accurate and informed clinical diagnoses. The classification of medical images is crucial to both medical management and educational endeavors [8]. The traditional approach's performance has reached its apex, though. In addition, it takes a lot of effort and time to extract and select classification parameters when using them [9]. The DCNN is an innovative approach to machine learning that has proven beneficial for a variety of categorization issues. CNN excel at a number of picture categorization tasks, producing the best results. However, medical image collections are difficult to compile since classifying them calls for an exceptionally high degree of professional competency [10].

Deep Convolutional Neural Networks (DCNNs) have also been employed for medical image fusion, offering several applications in healthcare. DCNNs can fuse low-resolution medical images with high-resolution images to generate enhanced, high-resolution images. This technique can be particularly beneficial in medical imaging, where higher resolution can provide better visualization of fine details, aiding in accurate diagnosis and treatment planning. DCNN-based fusion methods can fuse multiple images to generate a fused image with improved segmentation accuracy. By integrating information from different imaging modalities or perspectives, the fused image can provide more accurate and reliable boundaries and regions of interest for subsequent analysis and treatment planning [11].

DCNNs can be utilized for medical image registration, which involves aligning images from different modalities or time points. By fusing information from multiple images, DCNN-based methods can improve the accuracy and robustness of image registration, allowing for more precise analysis, monitoring, and treatment planning. It can fuse images from different sources to synthesize new images with desired characteristics or properties [12]. For example, fusing images from different imaging modalities can create a synthesized image that combines the strengths of each modality, providing comprehensive information for clinical analysis and decision-making. DCNNs can be employed for restoring medical images that are corrupted by noise, artifacts, or other degradations [13]. By fusing multiple degraded images, DCNN-based methods can effectively de-noise and enhance the image quality, enabling better visualization and interpretation of medical conditions. These applications demonstrate the versatility and effectiveness of DCNN-based medical image fusion techniques in improving image quality, accuracy, and clinical decision-making in various healthcare scenarios [14].

The key Contributions of this Research work is:

- The ensemble learning framework involves training multiple DCNN models independently on respective input modalities, capturing modality-specific information.
- A fusion module is designed to combine the extracted features from individual modalities, employing a weighted averaging technique to assign relevant and significant weights.
- The fused image obtained through this process exhibits improved spatial details, enhanced feature representation, and better preservation of diagnostic information.
- Thorough testing on a diverse dataset confirms the efficacy, visual quality, and robustness of the proposed method, showcasing its potential for broad adoption in clinical practice.

The manuscript of the approached paper is organized as follows: In Section II, some related works are reviewed. In Section III, Information regarding the problem statement is provided. In Section IV, the proposed Multi-Modal Image Fusion is covered in detail. In Section V, experiment results are provided, and discussed in Section VI with an extensive evaluation of the proposed approach to current best practices is made. In Section VII, the conclusion of the paper is provided.

II. Related Works

Maqsood et al. [15] suggested a multimodal fusion of images approach is based on limited representation and two-scale picture segmentation. The original heterogeneous images are initially subjected to contrast enrichment processing in the proposed system, which improves the brightness distribution for better visualization. The edge data gathered from intensity extended images is extracted using a spatial gradient-based edge detection method. The fundamental and detail layers are separated from the improved multiple mediums images at this point. Utilizing SSGSM, the final detailed layer is extracted. Finally, the fused image is produced utilizing an improved judgement maps and fusion scheme. By conducting both quantitative and qualitative evaluations, the experimental results demonstrate that the recommended multimodal picture fusion strategy outperforms several previous methods. However, it could happen for certain data from the initial images to be destroyed or distorted during the fusion process. The fusion mechanism may prioritize some qualities or aspects while ignoring others, resulting in the loss of crucial information or subtle traits.

Dinh et al.[16] proposed that the following are the key phases in the unique strategy that was presented to address the aforementioned shortcomings. In order to acquire the basic and detail elements, the three-scale deconstruction (TSD) approach is initially presented. Second, the output picture is fused using a rule based on the nearby energy function and the Kirsch compass operator, which aids in the retention of critical information. Thirdly, to fuse base layers with the best characteristics and produce a high-quality picture, the Marine Processors Algorithm (MPA) is used. This work compared the effectiveness of the suggested technique using six photograph quality criteria and five cutting-edge medical image fusion algorithms. Experiments revealed that the proposed method significantly increased the level of quality of the fusion picture and preserved edge information. The particular fusion algorithm used has a significant impact on the effectiveness of multidimensional picture fusion. Additionally, there doesn't exist a one-size-fits-all solution, and different techniques may yield different fusion results. The effectiveness and level of quality of the combined image can be considerably impacted by the algorithm choice.

Diwakar et al. [17] proposed a novel shearlet region multiple modalities image fusion method. The recommended technique uses Non-Subsampled Shearlet Transformation (NSST) to separate input pictures into low- and high-frequency parts. The localized extrema (LE) method is a unique technique used to separate and merge the fundamental layer and details layers. The co-occurring filter (CoF) is then used to combine the foundation layer and detail layer in harmonics with smaller elements. A high-frequency component is integrated using a sum modulated Laplacian (SML) as a component of an edge-preserving technique to image fusion. On the Multi-modal healthcare picture collection, experimental findings and contrasting assessment are performed using both recommended and modern methodologies. The recommended strategy beats cutting-edge fusion techniques in terms of blade retention in both objective and subjective assessment requirements, according to test findings and assessments. Numerous multidimensional merging of images algorithms is computationally demanding, requiring a significant amount of time and computing capacity. This could be a drawback in situations or real-time applications that need for quick fusion.

Stimpel et al. [18] demonstrated the globally linear guided filter for general medical image processing when coupled with a learning guiding map. The guided filter is the only element processing the output images, and its direction map may be trained to do the task optimally from beginning to end. The demising and graphic high-resolution tests are the two most often used activities when using this method to measure performance. The evaluation is based on cross-modal data sets that are paired. Modern methods are coupled with the provided procedure to achieve both goals. This can also show that the input image's information is basically unaltered after treatment, in contrast to conventional deep neural network approaches. The suggested pipeline also offers greater resilience against adversarial attacks and deteriorated input. Image fusion requires accurate registration of images from different modalities to align corresponding anatomical or functional structures. However, image registration can be challenging due to differences in acquisition protocols, patient motion, and anatomical variations. Registration errors can lead to misalignment and distortions in the fused image, affecting the accuracy of subsequent analysis.

Asha et al. [19] suggested a chaotic grey wolf optimization algorithm-based balanced blending of high-energy sub-bands of the Non-Subsampled Shearlet Transform (NSST) domain. The raw images are first dissected into their many scales and multi-directional components using the NSST. The modest number of pathways were combined according to a simple maximum rule in order to sustain the energy of an individual. In order to combine images of various frequencies and minimize the difference between the resultant image and the starting point pictures while retaining the textural characteristics of the input images, a collection of automatically adjusted high-frequency images is used. The major goals of the entire procedure are to maintain the energy of a low-frequency region while transferring textural details from the source images to the fused image. In order to construct the fused picture, the inverse NSST of the combining minimal and high-energy bands is used. Eight distinct illness datasets from Brain Atlas are used in the trials. More than 100 picture pairings are used to evaluate the efficacy of the suggested strategy using both objective and subjective quality evaluation. Due to the lack of contemporaneous collection of several modalities or the difficulty in gathering ground truth annotation for fusion quality, obtaining grounding truth for multipurpose fusion in medical imaging is problematic. Due to this, evaluating and comparing fusion procedures quantitatively is more difficult and frequently relies on opinions or substitute measurements.

Li et al. [20] To address the issue of poor contrast detail, a powerful image fusion technique employing numerous prominent features and a guided image filter was presented. The input photos were first divided into a number of calming and thorough images that had different scales before being subjected to the directed picture filter. Second, two different algorithms are used to extract important characteristics from the broken-down dependent upon visuals alongside the complete images in order to develop the combination rules. These two algorithms are the spectral residual (SR) technique for the mainframe gathering and the graph-based visually prominence model for a gradient saliency information extraction. The decomposition factors are combined using a process known as generalized intensity-hue-saturation (GIHS). The fused image is then reconstructed from the combined smoother and detailed images. The experimental findings show that, in the fields of MRI-PET and MRI-SPECT fusion, the proposed algorithm can outperform previous fusion approaches. The acceptability and use of fusion procedures in clinical practice, where openness and comprehensibility are vital, may be hampered by this lack of comprehension. The availability of information for the various modalities in multipurpose medical imaging may not be equal, meaning that one modality may contain greater numbers of specimens than the others. The fusion process may be impacted by this modality imbalance, which might result in biased fusion findings or a restricted representation of less common modalities.

Dai et al. [21] suggested that transformers have enormous promise for multimodal medical picture categorization. The proposed approach is based on the successful extraction of the link among sequences by the transformer. However, due to the small dimensions of medical information sets for pictures and the lack of sufficient data to establish the connection between low-level semantic variables, the precision of pure transformation systems based on ViT and DeiT is not good in versatile classification of medical images. TransMed is therefore suggested as a way to collect both cross-modality high-level information and low-level characteristics.

TransMed combines the benefits of both CNN and Transformer. TransMed converts the multimodal pictures into sequences, delivers them to CNN for processing, and then use transformers to discover the connections between each sequence and provide predictions. TransMed beats the current multipurpose fusion approaches when it comes to of parameters, operating speed, and accuracy because the transformer successfully models the global aspects of multifaceted pictures. Finding the best fusion approach, though, is a challenging task. Different fusion methods, each with various advantages and disadvantages, may be used, including pixel-level, decision-level fusion and feature-level. For a certain application or modality combination, choosing the best fusion approach necessitates extensive thought and skill.

III. Problem Statement

Multi-modal medical imaging provides valuable complementary information for accurate diagnosis, treatment planning, and monitoring of various diseases. The issue of successfully integrating and fusing data from many imaging modalities is still difficult. Traditional fusion methods' dependence on ad hoc extraction of features and fusion techniques that commonly use handmade feature extraction that approximate complicated interactions between paradigms may restrict the quality of the merged image. Furthermore, the actual applicability of these approaches in clinical contexts is hampered by their lack of stability and interpretability. CNN have proven to perform exceptionally well in a variety of computer vision applications, including the processing of medical images. CNNs have not yet been extensively used in multi-modal medical picture fusion, nevertheless. The research gap in the mentioned existing works lies in the need for more comprehensive and adaptable multimodal image fusion techniques that can simultaneously address various aspects of quality enhancement, edge preservation, and overall visual fidelity. The proposed DCNN aims to overcome the limitations of traditional fusion methods and single CNN-based approaches by effectively capturing the complementary information present in multiple modalities and improving the fusion quality. The ensemble learning framework is expected to leverage the diversity and strengths of individual networks to enhance the accuracy, robustness, and interpretability of the fusion process [22].

IV. Proposed Ensemble Learning Approach

The suggested method entails enhancing the supplied image. Then DCNN are used to accomplish Medical Image Fusion. The performance is then assessed. The suggested method for Multi-Modal Medical Image Fusion using DCNN is shown in Fig. 1. The input photographs are first preprocessed by converting them to a standard scale and using the proper transformations to improve image details. Then, using a sizable dataset of aligned multi-modal pictures and a fusion-specific loss function, a CNN architecture is created, consisting of shared and modality-specific convolutional layers. From each modality, high-level feature maps are retrieved using the trained CNN.

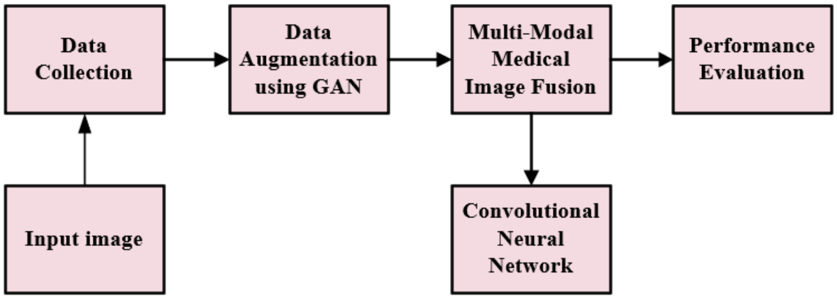
Data Collection
Data Augmentation using GAN
Multi-Modal Medical Image Fusion
Performance Evaluation
Input image
Convolutional Neural Network

Fig. 1. Proposed approach for multi-modal medical image fusion.

A. Data Collection

The BRATS 2018 dataset, which contains Multi-Modal MRI images and patients' healthcare information with distinct heterogeneous histologic sub-regions, different levels of aggression, and variable prognosis, was used for training and testing in this work. These clinical multi-modal MR images have been generated using a range of magnetic field intensities and scanners [23]. Table I describe the dataset for Training and Validation.

TABLE I. The Collected Datasets

	Training data	Testing data
Unhealthy	350	350
Healthy	250	250
Overall data	600	600

B. Data Augmentation using GAN

Data analysis, enhancement, combination, and rescaling are all part of preprocessing. The acquired source photographs are transformed to RGB images before augmentation. The improvement method is used to build more powerful simpler models that are impervious to some sorts of picture manipulation, following which the image's quality is altered to improve the information's integrity and degree of variability. The layering placement of the original photos is important for the concatenation. The first phase is the R channel of an MRI, and the second layer is the R channel of a PET scan (positron emission tomography). In the second layer, the B(Blue) channels of MRI and the B channels of PET are placed after the G(Green) channels of MRI and the G channel of PET, respectively. The pictures that offer practical details must be maintained below the photos that offer structural details in all pathways, it must be highlighted [24].

Fig. 2 the generator creates a picture according to the parameters that are collected from the image, based on the number of channels supplied in the layers for input and output. The modelled output picture of the suggested method contains three channels and six input channels. The generator automatically fits the obtained parameters into the three stated channels during training. The stacked order is set to RR1GG1BB1 and the training data is compressed to prevent the color space from being disrupted. The RGB components of the first source picture are R, G, and B, while the RGB components of the additional source image are R1G1B1. Random switching and unpredictability are employed for data augmentation. With random flipping, there is a 50% chance that the picture will be turned. The picture is accompanied with random noise, which is Gaussian in nature with an average value of 0 and a variation of 0.1. This method can be used to learn the aggregate breakdown of single-modality imaging information as well as for recording the broad distribution of imaging data from several modalities. The primary producer can learn to produce many modalities at once since different modalities' information collected from a single ROI share identical information with unique appearance patterns. Such a generator can be used to complete missing modes of operation or supplement data [25].

As an estimate $K_{data}(u)$, GAN aims to learn an estimate of probabilities, $k_G(u)$, from the actual distribution. u= G(v), the sample, where the noise variable is called v. It resolves the issue by simultaneously instructing the generator N and a discriminator D to create a process that is adversarial. By sampling noise, G produces samples from latent space. Whether the sample comes from $K_G(u)$ or $K_{data}(u)$ is determined by D. G samples eventually approach genuine or real samples through the continuous unfavorable effect. The definition of the optimization formula D is represented in Eq. (1) [26].

$$D^* = argminDiv_D\big(k_D(u), k_{data}(u)\big) \quad (1)$$

Where Div (*) indicates the divergence among the two distributions. N may be used to compute the divergence and generate the following objective function as represented in Eq. (2)

$$N^* = argmaxV_F(G, D) \quad (2)$$

Where,

$$V(G, D) = J_{X\sim Kdata}\big[\mathrm{logD}(x) + j_{x\sim kG}\big[\mathrm{log}\big(1 - D(u)\big)\big]\big] (3)$$

Hence, the Eq. (1) is transformed as

$$G^* = argmin_N max_D V(N, D) \quad (4)$$

In contrast to a traditional GAN, which consists of a single generator and a discriminator, pix2pixHD uses an auxiliary producer and a primary generator to output pictures at two distinct resolutions, which are 3x448x448 and 3x224x224 in this instance. Therefore, two entirely convolutional network-based discrimination named D_p and D_q are in charge of the two solutions.

C. *Multi-Modal Fusion-CNN*

Patch incorporation, class insertion, position integration, class token and patch token are the five insertions and tokens that are present in the input layer. While class anchoring is an adaptable vector, patch anchoring represents each patches' input from CNN. Using position embedded data and patched embedded data; this technique preserves the geographical and geographical data of a patch by encoding it into patch tokens. Class signaling and class anchoring are equal since category anchoring does not provide patch embedding. The Eq. (5) and (6) represents the input is u, the adaptable vector is V_a, the location embedding is u_{pa}, the patch tokens are u_{pq}, and the class token is u_{de}.

$$u_{pq} = Conv(u) + u_{pa} \quad (5)$$

$$u_{de} = V^a \quad (6)$$

The type token connects to the patched tokens preceding the converters' input layer, goes via the conversion layer, and is subsequently generated from the fully connected layer in order to foresee the class. The core of the arrangement is an image power source, which receives images from various input modalities and generates a task-optimal unified depiction of the required guiding map. In extracting the most important information directly from data, convolutional neural networks (CNN) have demonstrated significant success. A CNN is applied to build the guiding map as a result. De-noising and picture super resolution are the two tasks we focus on. The guide maps for both are generated using tested network designs for the sake of repeatability. The necessity to handle numerous input photos led to the sole adjustments. The inclusion of more guiding photos would logically be conceivable and is only constrained by availability and processing capacity. In order to determine how the selected network design affects the guided filtering process, employs using two separate networks for super resolution [27].

Fig. 3 represents the CNN architecture where a layer of neurons is fully linked, every neuron in that layer is also connected to every neuron in the layer underneath it. The value should indicate the degree to which of the connection between the neuron that is j^{th} in this particular stratum and the kth neurons in the preceding layer $\partial_{0_{lk}}$. Let b_{1_j} be the bias of the jth neuron in the current layer. The result of the layer's j^{th} neuron is given by Eq. (7).

$$y_{0_j} = \sum_k \partial_{1_{lk}} x_{0_k} + b_{1_j} \quad (7)$$

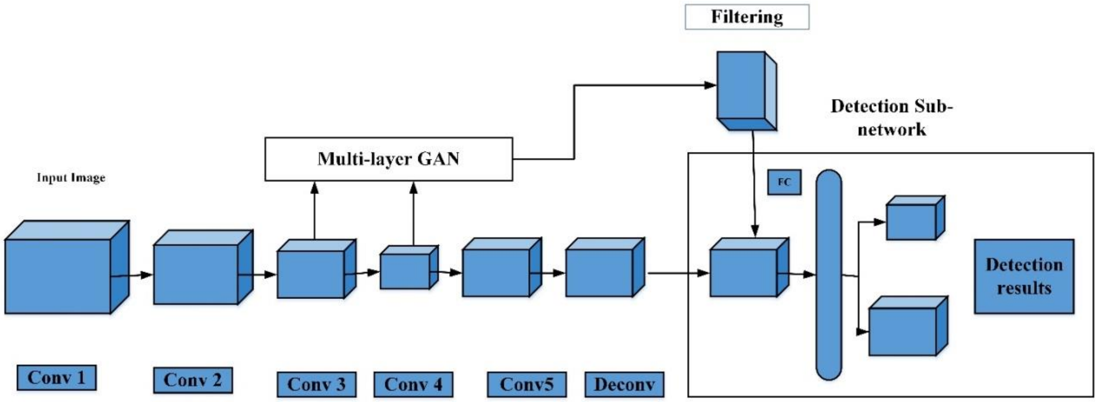

Fig. 2. GAN in data augmentation.

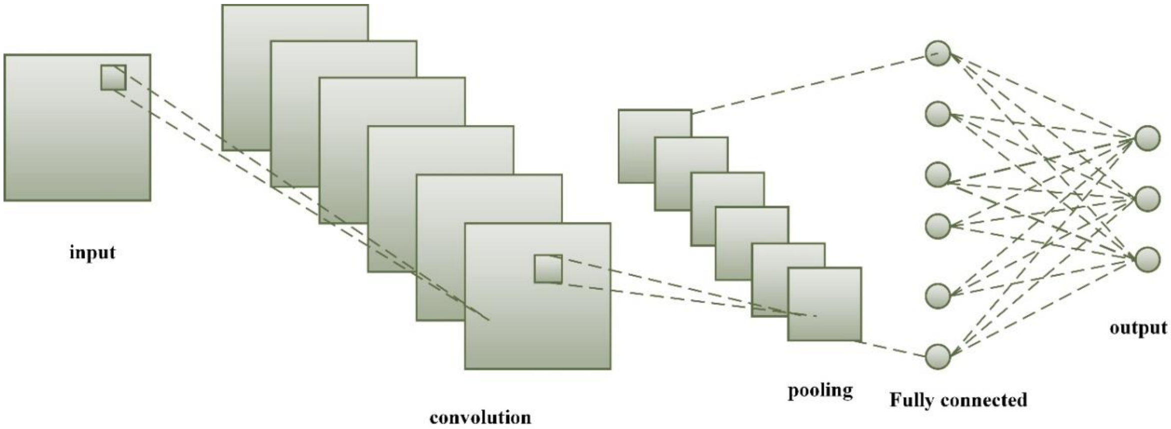

Fig. 3. CNN architecture.

The convolutional layer's neurons that are frequently utilized to produce a kernel or filtration have the same biases and values. If the dimensions of the filtering are set to nxn, every neuron in the corresponding layer will be linked to a nxn area of the neurons that are in the layer above it. In line with this, the (j, k)th neuron's outputs will be in Eq. (8)

$$y_{1_{j,k}} = \sum_{p=0}^{n-1} \sum_{q=0}^{n-1} \omega_{0_{p,q}} x_{0_{j+p,k+q}} + c \qquad (8)$$

Examples of regularly used activation functions include Tanh, Sigmoid, and the inverted linear unit, which is now the de facto recommendation for contemporary neural networks. Convolutional or completely linked layers are typically followed by activation layers to provide elementwise non-linear behavior. By using the activation function, is defined in Eq. (9)

$$f_0(x) = \max(x, 0) \qquad (9)$$

The downwards sampling method for every sub-area in the pooling layer provides the dimension of a single neuron in the present one by dividing the neurons of the layer preceding it into an array of not overlapping rectangles. Maximum pooling and average-pooling, the two most popular pooling procedures, offer the subarea's maximum value and average value, respectively. A convolutional neural network usually sets up a sequence of convolutional (Conv)-ReLU layers, before adding the pooling layers (Pool), and continues doing this till the picture gets spatially combined to a compact size. At certain points, it is usual to switch to fully-connected layers (FC). Three different parts make up each neuron: the receiving domain, a modulation domain, and a pulse-generating domain. Feedback is created by the connections between many neurons. An external input signal is first received in the receiving domain, after which it is amplified in the modulating domain and the final output pulse is produced in the pulse generating domain. The fundamental procedure is as follows: signals from the feedback channel domain and the link domain are received in the receiving domain, and they travel through Channels L and F into the modulation domain.

The necessary feature pixels in the layer of convolution are added to each image's output pixel after synchronizing the characteristics from the source pictures. Add every value of a pixel together, and then divide the result by the total number of pixels in the description. The feature map has been added to the computed values, causing the improvement to be applied to the whole image. The characteristics map has a slot for each computed value. All of the traits are therefore processed, and several feature maps are produced. The Eq. (10) to obtain the convolutional layer is the following,

$$v_{xyz} = \sum_{E=0}^{E-1} \sum_{F=0}^{F-1} \sum_{H=0}^{G-1} s_{x+g,y+g,E^e eghe}^{(l-1)} + f_{pqr} \qquad (10)$$

Where f_{pqr} is generally set to which is not contingent on the image's component position. $E^e eghe$ as an identical value of weight. Since it recovers the distinguishing properties of the image using various convolution kernel sizes, the layer of convolution is a critical part of CNN. The layers of inversion can be continuously applied to the input photos to create a set of feature maps. K_i may then be created by using the characteristic map of the i-th layer in CNN as it is represented in Eq. (11)

$$K_i = \rho(K_{i-1}V_i + H_i) \tag{11}$$

Where K_i is the current networks layer's mapping of features, D_{i-1} is the previous layer's convolution feature. V_i is the i-th layer weight, k_i is the i-th layer offset vector, and ρ (·) represents the rectified function. Layer pooling's goal is to reduce the total amount of space, that can cut processing costs and effectively reduce the danger of over-fitting. The resultant characteristic on the ith localized responsive field is determined in Eq. (12) in the k-th layer of pooling.

$$u_i^k = down(u_i^{k-1}, s) \tag{12}$$

where down (·) indicates the function for down-sampling, u_i^{k-1} is the feature vector in the previous layer, and r is the pooling size. Following the pooling and convolutional layers, there may be one or more fully-connected (FC) layers, which use the collected features for picture categorization. It classifies the input brain images into healthy and unhealthy.

D. *Multi-Modal Fusion*

In order to process sources of any size, the conversion phase is employed throughout the picture checking and fusion procedure on the totally connected layer. Using the same kernel size, the entire connected layer is split into two comparable convolutional layers. The network may then combine pictures of any size, X and Y, to produce a dense prediction map, I. Each prediction Is on the map is represented by a vector with two dimensions with values between 0 and 1. If one dimension of a prediction is bigger than the other, it is normalized to 1 while the other dimension is set to 0, making the weights given to related image blocks easier. With an outcome aspect value of 1, this ensures that the weight of every image block is decreased. Two near forecasts in S have overlapping areas in their corresponding picture blocks. The weights of the photos in these overlapped portions are added to determine the mean value of the adjacent picture blocks. The network may be given pictures of any size, both X and Y, using this technique, and a weight map W of the same size is generated. This guarantees a weight reduction for each picture block with an output aspect value of 1.

E. *Fusion Rules*

In order to attain better look, richer details, and spectacular fusion impacts, this study suggests novel fusion principles and the average weighted fusion operations in accordance with area peculiarities. The fusion guidelines and commands are as follows:

Stage 1: It determines the energy $R_u^o \, and \, R_v^o$ of matching localized areas in each breakdown layer o of source images x and y, accordingly, using the contrast pyramid deconstruction:

$$R_u^o(a,b) = \sum_x \sum_y S_u^o(a+u, b+m)^2 \tag{13}$$

$$R_u^o(a,b) = \sum_x \sum_y S_v^o(a+u, b+m)^2 \tag{14}$$

Where Equations (13) and (14) the regional area power $R_u^o(a,b)$ on the oth layer of difference is centered at (a, b). structure, where u and v stand for the size of the region in question, and represents the image of the contrast between the structuring fourth layer.

Stage 2: Determine how similar the respective local areas in two source photos are to one another.

Stage 3: Decide who the fusion operators.

As a consequence, the strategy selects the center pixel based on energy variations when the degree of similarity is below the threshold of significance and employs the weighted fusion operator when it is equal to or above.

Algorithm 1: Multi-Modal Medical Image Fusion using Deep Convolutional Neural Networks

```
Input: Medical Images
Output:   fusion result
The two source images and the initial fused one are given
Train the input images vi in the system, where i = 1 to n

Data Augmentation of images
Let U(i) be the input images from the dataset                    // using GAN
            for every  $U_i$
                        $V_v(i) = V(i) – N$                              // V denotes unwanted noise

Segmentation of images
    Initialize the starting point of  the highlighted portion
    if (image detected)
           Gather the subset
           Identify the highlights in the hyperspectral image using Eq.  (7)
    Else
    Repeat until the stopping condition is reached              // until the image is identified
    End if
Return
Image Fusion using CNN
```

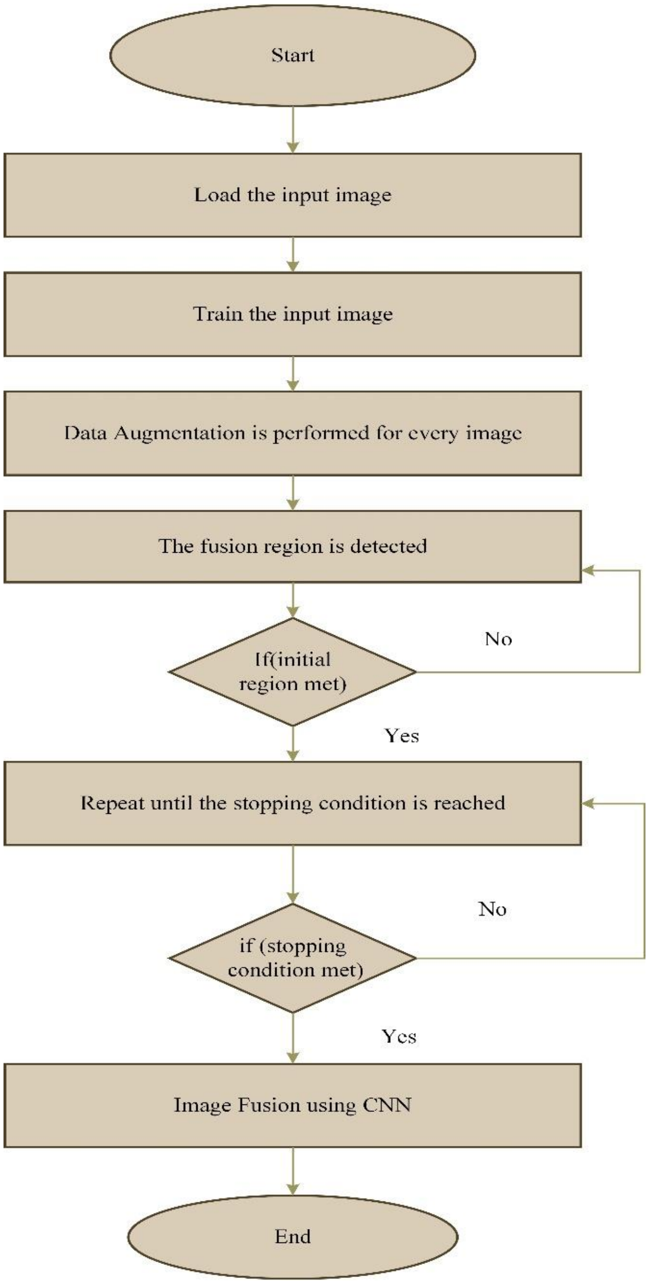
Start
Load the input image
Train the input image
Data Augmentation is performed for every image
The fusion region is detected
If(initial region met)
No
Yes
Repeat until the stopping condition is reached
if (stopping condition met)
No
Yes
Image Fusion using CNN
End

Fig. 4. Flow chart of the proposed system.

Fig. 4 represents the Ensemble Learning Approach for Multi-Modal Medical Image Fusion using Deep Convolutional Neural Networks.

V. Results

The recommended method has been evaluated using datasets and executed in MATLAB software on the Windows 10 platform. In order to solve this issue, deep CNNs are utilized in the article to extract high-level characteristics from the data modalities and NMF is employed to discover the fused image's underlying structure. The use of deep CNNs, which have demonstrated extraordinary capacity in understanding intricate patterns and characteristics from pictures, is a key benefit of the suggested technique. The model can accurately capture and reflect the unique qualities of each modality by utilizing the power of deep learning, thereby enabling a more thorough synthesis of information. In order to evaluate the effectiveness of their strategy, the authors additionally offer quantitative assessment criteria including precision, recall, precision, accuracy, F-score, specificity, and sensitivity. The suggested approach's robustness and dependability are highlighted by the excellent scores in these criteria that were attained. Overall, multi-modal image fusion using deep CNNs and NMF makes a significant addition to the discipline. The suggested approach successfully combines deep learning for feature extraction with NMF to train the fused representation, producing better fusion results. The results of the investigation and analyses show how this approach can be used for a range of applications, including mapping, and imaging in medicine. The use of multipurpose image fusion techniques in the health care imaging field is essential for better medical diagnosis and therapy. The research study suggests a unique method for fusing multimodal medical images that incorporates deep convolutional neural networks (CNNs).

A. *Accuracy*

The model's total Accuracy shows how well it performs across all classifications. In overall, it is the idea that every circumstance can be forecast with accuracy. Eq. (15) represents the Accuracy:

$$A = \frac{T_{pos}+T_{neg}}{T_{pos}+T_{neg}+F_{pos}+F_{neg}} \tag{15}$$

B. *Precision*

Precision is calculated as the total amount of positive predictions multiplied by the number of correct positive estimations. It measures how many accurately merged multi-modal medical pictures there are. Eq. (16), which is used to compute the accuracy

$$P = \frac{T_{pos}}{T_{pos}+F_{pos}} \tag{16}$$

C. *Recall*

The ratio of correct positive forecasts to true positives and false negatives is known as recall. It displays the proportion of correctly predicted events and picture fusion across different modes. The recall is represented by Eq. (17),

$$R = \frac{T_{pos}}{T_{pos}+F_{neg}} \tag{17}$$

D. *F1-Score*

Precision and recall are combined in the F1-Score calculation. The F1-Score as shown in Eq. 18) is created using precision and recall.

$$F = \frac{2\times Precision\times recall}{Precision\times recall} \tag{18}$$

E. *Sensitivity*

It is a measure of the proportion of correctly foretold true positives. Eq. (19) is used to calculate sensitivity as,

$$Sensitivity = \frac{T_{pos}}{T_{pos}+T_{neg}} \tag{19}$$

F. *Specificity*

The degree gauges identify precisely the true negatives. Eq. (20) is used to calculate the specificity value as,

$$Specificity = \frac{T_{neg}}{F_{pos}+T_{neg}} \quad (20)$$

TABLE II. COMPARISON OF ACCURACY

Classifier	Accuracy
CNN [10]	86.8
RNN [11]	97.9
KNN [14]	98.2
AlexNet[16]	98.5
DCNN	99.6

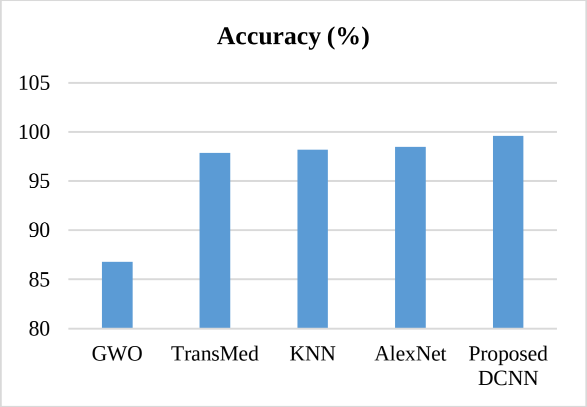

Fig. 5. Comparison of accuracy.

When compared to the current techniques, the suggested technique DCNN obtains a greater level of accuracy. The contrast of efficiency between DCNN and other approaches is shown in Table II and diagrammed in Fig. 5.

TABLE III. COMPARISON OF PRECISION AND RECALL

Methods	Precision (%)	Recall (%)
KNN	89.5	89.1
CNN	96.9	98.5
GWO	97.9	95
DCNN	99.9	99

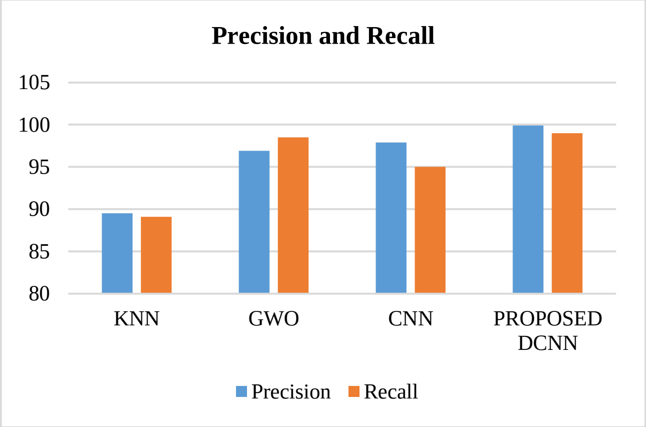

Fig. 6. Comparison of precision and recall.

Table III demonstrates that the proposed technique of combined DCNN achieves higher precision and recall of 99.9% and 99% when compared to the existing methods. The advanced DCNN gives better accuracy than the performance evaluated. Here, the achieved accuracy level is 99 using the DCNN model. Fig. 6 illustrates the precision and recall between DCNN and other methods**.** The model's balanced and trustworthy performance is further supported by the F-score which takes precision and recall into account. These findings support the suggested model's exceptional qualities, including precision, recall, precision, F-score, sensitivity, and specificity, which make it a trustworthy and efficient option for the task at issue.

TABLE IV. SENSITIVITY AND SPECIFICITY FOR PROPOSED METHOD

Proposed Model	
Sensitivity	98.14
Specificity	96.68

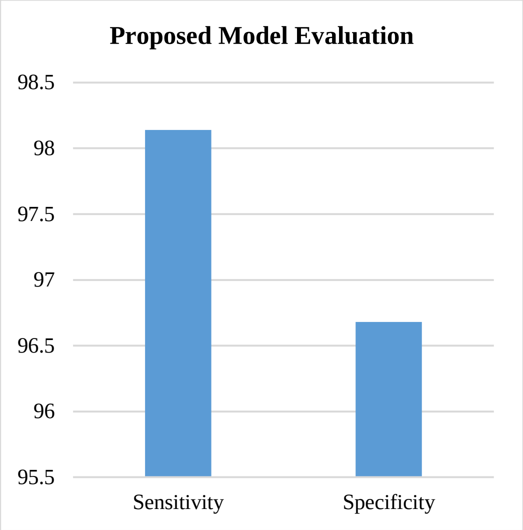

Fig. 7. Comparison of specificity and sensitivity.

Fig. 7 and Table IV represents the model's specificity score of 98.14% shows that there is little chance of making false positives for negative predictions, while its sensitivity score of 96.68% emphasizes how well the model can recognize positive circumstances.

The accuracy of the convolutional neural network used for both the training and testing stages is 99.4% and 97.5%, respectively, according to Table V. When DCNN is utilized, the accuracy of the testing and training processes increases to 99.9% and 99.4%, respectively. Fig. 8 shows an evaluation of performance.

TABLE V. PERFORMANCE EVALUATION

	CNN	ABO-CNN
Training	98.1	99.9
Testing	97.5	99.4

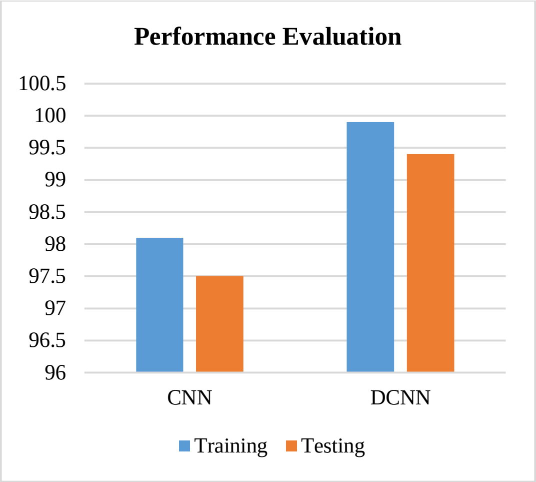

Fig. 8. Performance evaluation.

Table VI and Fig. 9 presents a comparison of medical image fusion techniques based on three evaluation metrics: Gradient-based quality, Information ratio, and Mutual information. Each metric is accompanied by corresponding percentages representing the performance of the techniques in relation to that metric. The Gradient-based quality metric is evaluated at 89%, 45.5%, and 67.7% for RMSE, indicating the percentage of quality achieved by the fusion techniques in terms of gradient-based measures. Similarly, the PSNR metric indicates a performance of 54%, 40%, and 79% for the techniques, representing the Peak Signal-to-Noise Ratio achieved by the fusion results. Lastly, the ASR metric is reported at 45%, 39.5%, and 59%, representing the Accuracy Success Rate of the fusion techniques. This table allows for a comparative analysis of different medical image fusion methods based on multiple evaluation metrics, providing insights into their respective performance levels across various quality measures.

TABLE VI. MEDICAL IMAGE FUSION COMPARISON

	Gradient-based quality	**Information ratio**	**Mutual information**
RMSE	89%	45.5%	67.7%
PSNR	54%	40%	79%
ASR	45%	39.5%	59%

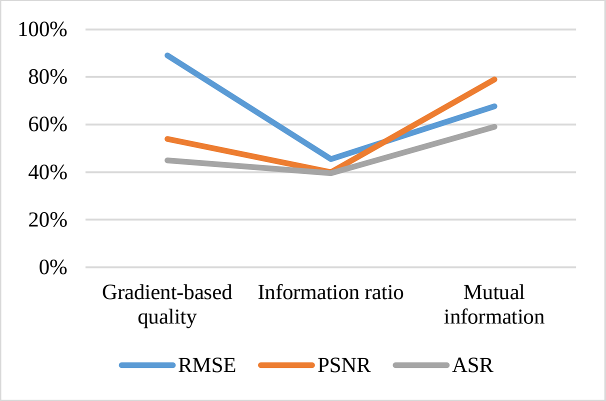

Fig. 9. Medical image fusion comparison.

TABLE VII. COMPARISON OF PROCESSING TIME

Methods	**Processing Time**
KNN	11.05
CNN	14.58
GWO	12.86
DCNN	6.15

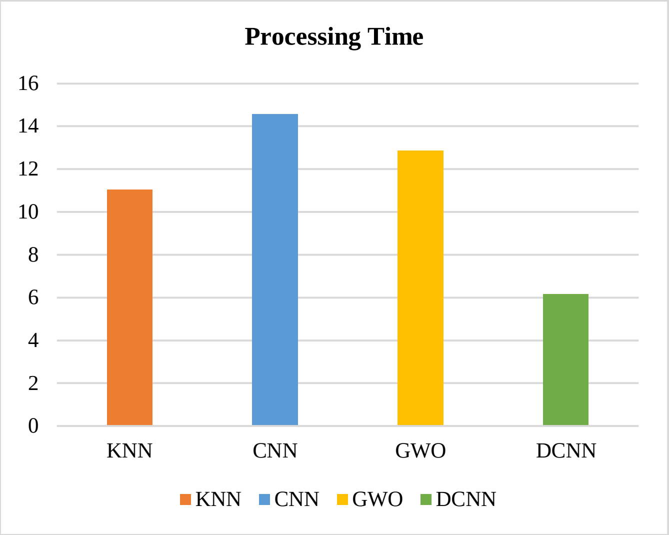

Fig. 10. Evaluation Comparison of processing time.

Table VII and Fig. 10 presents a comparison of processing times for different methods, namely KNN, CNN, GWO, and DCNN. The Processing Time column indicates the time taken by each method for a specific task or process. From the table, it can be observed that KNN takes 11.05 units of time, CNN takes 14.58 units, GWO takes 12.86 units, and DCNN takes 6.15 units. These values reflect the computational efficiency or speed of each method, with a lower processing time indicating faster execution. The table provides insights into the relative performance of these methods in terms of processing time, which can be valuable for selecting an appropriate method based on time constraints or efficiency requirements.

G. ROC Curve

Fig. 11 represents the ROC Curve of the proposed system. The proposed DCNN has the higher rate when compared to the existing methods. The ROC curve is a graphical representation of the performance of a binary classification system as its discrimination threshold is varied. However, the ROC curve is not directly applicable to evaluate multi-modal image fusion, as it is typically used for evaluating classification models.

H. Accuracy and Loss for Training and Validation

Fig. 12 represents the accuracy of a multi-modal image fusion model refers to how well it can effectively integrate and preserve relevant information from the input images while suppressing noise, artifacts, and inconsistencies.

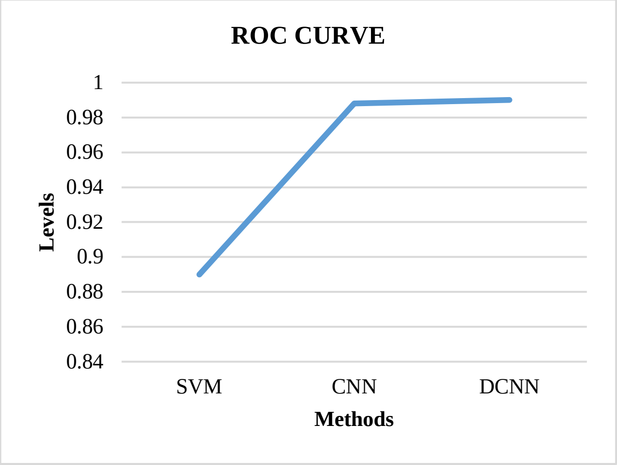

Fig. 11. ROC curve.

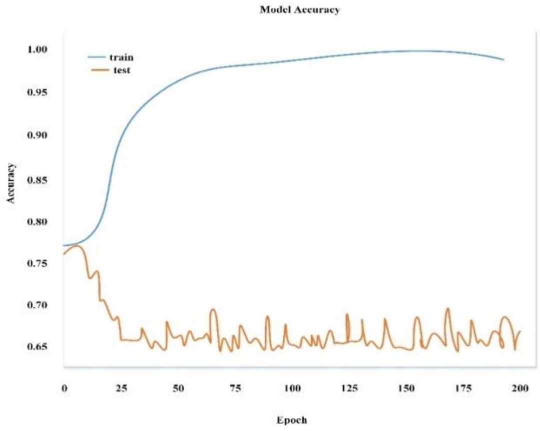

Fig. 12. Model accuracy for training and validation.

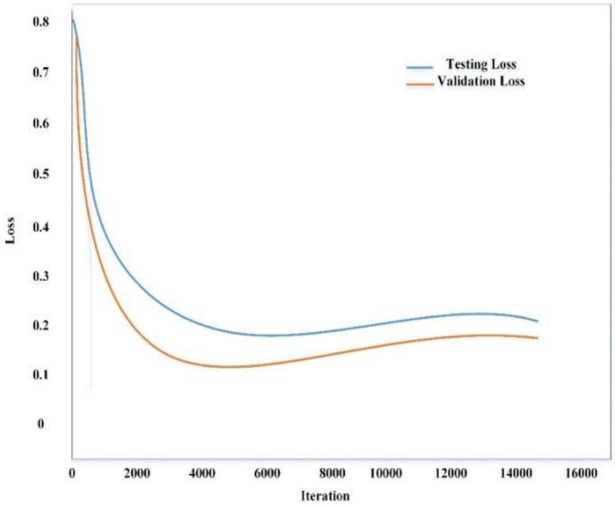

Fig. 13. Model loss for training and validation.

Fig. 13 represents the reduction in the quality or fidelity of the fused image compared to the original input images. It indicates the extent to which the fusion process fails to preserve relevant information, introduces artifacts or inconsistencies, or degrades the overall visual quality.

VI. Discussion

Existing techniques frequently concentrate on certain aspects of image fusion, such as feature extraction, detail preservation, or computational effectiveness, but there are no complete solutions that handle all elements of quality, such as contrast augmentation, edge preservation, and overall visual fidelity. The performance of fusion approaches across diverse modalities, clinical applications, and data quantities cannot be fully evaluated due to the lack of defined assessment parameters. It is still difficult to get timely collections of multi-modal data and trustworthy annotations, hence the problem of ground truth annotation for quality evaluation in medical image fusion persists [28]. Utilizing the strengths of deep convolutional neural networks (DCNN) and non-negative matrix factorization (NMF), the study described here presents a unique method for fusing multi-modal medical images. Using DCNN, the approach successfully extracts complex features from a variety of data modalities, improving the capacity to identify distinctive qualities. Applying NMF next reveals the fused image's underlying structure. Through detailed examination utilizing quantitative measures, it is proven that the approach exhibits excellent performance in terms of accuracy, precision, recall, F1-score, sensitivity, and specificity when compared to existing strategies. Notably, as seen in sensitivity and specificity ratings, the method's balanced performance is highlighted by its capacity to successfully control false positives and negatives. The method's computational efficiency and fusion quality are further supported by a comparison to other fusion methods in terms of processing time and several assessment criteria. Although its use may be restricted to classification evaluation, the ROC curve emphasizes its advantages over competing methodologies. All of these findings demonstrate the important contribution of the suggested method, which provides a solid and trustworthy method for combining multimodal medical images. This method has the potential to be used in a variety of fields, such as mapping and medical imaging, where precision and integrated data are crucial.

VII. Conclusion

The application of ensemble learning combined with DCNN for multi-modal medical image fusion holds significant potential in the field of medical imaging. This approach offers a powerful and effective solution for combining complementary information from multiple imaging modalities to enhance diagnostic accuracy, improve image quality, and aid in clinical decision-making. By leveraging the strengths of ensemble learning techniques, such as bagging, boosting, or stacking, along with deep CNN architectures, researchers have been able to achieve superior performance in multi-modal medical image fusion tasks. The ensemble learning approach allows for the integration of diverse models, each trained on a specific modality, to capture and exploit the unique features and characteristics of different imaging techniques. Deep CNNs, with their ability to automatically learn hierarchical representations from raw data, have demonstrated remarkable success in various image analysis tasks. They provide a suitable framework for effectively extracting relevant features from multi-modal medical images and fusing them to generate a fused image that preserves crucial information from each modality. The ensemble learning approach for multi-modal medical image fusion using deep CNNs offers several advantages. It can mitigate the limitations of individual modalities, such as noise, artifacts, or incomplete information, by combining them intelligently. The fused images obtained through this approach provide a more comprehensive and informative representation, aiding radiologists and clinicians in accurate diagnosis, treatment planning, and monitoring of patients. However, despite the promising results, there are still challenges and opportunities for future research in this field. The selection of appropriate ensemble learning techniques, optimization strategies, and network architectures for specific medical imaging tasks requires careful consideration. Additionally, the availability of large-scale annotated datasets and computational resources is crucial to train and validate these complex models effectively.

REFERENCES

[1] M. Wei, M. Xi, Y. Li, M. Liang, and G. Wang, “Multimodal Medical Image Fusion: The Perspective of Deep Learning,” Academic Journal of Science and Technology, vol. 5, no. 3, Art. no. 3, May 2023, doi: 10.54097/ajst.v5i3.8013.

[2] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, “Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI,” Information Fusion, vol. 71, pp. 28–37, 2021.

[3] A. Rossi, M. Hosseinzadeh, M. Bianchini, F. Scarselli, and H. Huisman, “Multi-Modal Siamese Network for Diagnostically Similar Lesion Retrieval in Prostate MRI,” IEEE Transactions on Medical Imaging, vol. 40, no. 3, pp. 986–995, Mar. 2021, doi: 10.1109/TMI.2020.3043641.

[4] Y. Cao, L. Cui, L. Zhang, F. Yu, Z. Li, and Y. Xu, “MMTN: Multi-Modal Memory Transformer Network for Image-Report Consistent Medical Report Generation,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, Art. no. 1, Jun. 2023, doi: 10.1609/aaai.v37i1.25100.

[5] C. Jiang, Y. Chen, J. Chang, M. Feng, R. Wang, and J. Yao, “Fusion of medical imaging and electronic health records with attention and multi-head machanisms.” arXiv, Dec. 22, 2021. doi: 10.48550/arXiv.2112.11710.

[6] M. Haribabu, V. Guruviah, and P. Yogarajah, “Recent Advancements in Multimodal Medical Image Fusion Techniques for Better Diagnosis: An Overview,” Current Medical Imaging Reviews, vol. 19, no. 7, pp. 673–694, Jun. 2023, doi: 10.2174/1573405618666220606161137.

[7] A. Pemasiri, K. Nguyen, S. Sridharan, and C. Fookes, “Multi-modal semantic image segmentation,” Computer Vision and Image Understanding, vol. 202, p. 103085, Jan. 2021, doi: 10.1016/j.cviu.2020.103085.

[8] V. S. Parvathy and S. Pothiraj, “Multi-modality medical image fusion using hybridization of binary crow search optimization,” Health Care Manag Sci, vol. 23, no. 4, pp. 661–669, Dec. 2020, doi: 10.1007/s10729-019-09492-2.

[9] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, and X. Zhang, “Multi-modal deep learning model for auxiliary diagnosis of Alzheimer’s disease,” Neurocomputing, vol. 361, pp. 185–195, Oct. 2019, doi: 10.1016/j.neucom.2019.04.093.

[10] J. Zhang et al., “Joint Vessel Segmentation and Deformable Registration on Multi-Modal Retinal Images Based on Style Transfer,” in 2019 IEEE International Conference on Image Processing (ICIP), Sep. 2019, pp. 839–843. doi: 10.1109/ICIP.2019.8802932.

[11] D. Kumar and D. Sharma, “Multi-modal Information Extraction and Fusion with Convolutional Neural Networks,” in 2020 International Joint Conference on Neural Networks (IJCNN), Jul. 2020, pp. 1–9. doi: 10.1109/IJCNN48605.2020.9206803.

[12] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, “Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training,” IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 12, pp. 6070–6080, Dec. 2022, doi: 10.1109/JBHI.2022.3207502.

[13] D. Mussina, A. Irmanova, P. K. Jamwal, and M. Bagheri, “Multi-Modal Data Fusion Using Deep Neural Network for Condition Monitoring of High Voltage Insulator,” IEEE Access, vol. 8, pp. 184486–184496, 2020, doi: 10.1109/ACCESS.2020.3027825.

[14] C. Cui et al., “Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review,” Prog. Biomed. Eng., vol. 5, no. 2, p. 022001, Apr. 2023, doi: 10.1088/2516-1091/acc2fe.

[15] S. Maqsood and U. Javed, “Multi-modal Medical Image Fusion based on Two-scale Image Decomposition and Sparse Representation,” Biomedical Signal Processing and Control, vol. 57, p. 101810, Mar. 2020, doi: 10.1016/j.bspc.2019.101810.

[16] P.-H. Dinh, “A novel approach based on Three-scale image decomposition and Marine predators algorithm for multi-modal medical image fusion,” Biomedical Signal Processing and Control, vol. 67, p. 102536, May 2021, doi: 10.1016/j.bspc.2021.102536.

[17] M. Diwakar, P. Singh, and A. Shankar, “Multi-modal medical image fusion framework using co-occurrence filter and local extrema in NSST domain,” Biomedical Signal Processing and Control, vol. 68, p. 102788, Jul. 2021, doi: 10.1016/j.bspc.2021.102788.

[18] B. Stimpel, C. Syben, F. Schirrmacher, P. Hoelter, A. Dörfler, and A. Maier, “Multi-Modal Deep Guided Filtering for Comprehensible Medical Image Processing,” IEEE Transactions on Medical Imaging, vol. 39, no. 5, pp. 1703–1711, May 2020, doi: 10.1109/TMI.2019.2955184.

[19] C. S. Asha, S. Lal, V. P. Gurupur, and P. U. P. Saxena, “Multi-Modal Medical Image Fusion with Adaptive Weighted Combination of NSST Bands Using Chaotic Grey Wolf Optimization,” IEEE Access, vol. 7, pp. 40782–40796, 2019, doi: 10.1109/ACCESS.2019.2908076.

[20] W. Li, L. Jia, and J. Du, “Multi-Modal Sensor Medical Image Fusion Based on Multiple Salient Features with Guided Image Filter,” IEEE Access, vol. 7, pp. 173019–173033, 2019, doi: 10.1109/ACCESS.2019.2953786.

[21] Y. Dai, Y. Gao, and F. Liu, “TransMed: Transformers Advance Multi-Modal Medical Image Classification,” Diagnostics, vol. 11, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/diagnostics11081384.

[22] F. Xiao, B. Li, Y. Peng, C. Cao, K. Hu, and X. Gao, “Multi-Modal Weights Sharing and Hierarchical Feature Fusion for RGBD Salient Object Detection,” IEEE Access, vol. 8, pp. 26602–26611, 2020, doi: 10.1109/ACCESS.2020.2971509.

[23] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, “Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images,” Scientific Reports, vol. 11, no. 1, p. 10930, 2021.

[24] Q. Chang et al., “Multi-modal AsynDGAN: Learn from Distributed Medical Image Data without Sharing Private Information.” arXiv, Dec. 15, 2020. doi: 10.48550/arXiv.2012.08604.

[25] Z. Chen, J. Wei, and R. Li, “Unsupervised Multi-Modal Medical Image Registration via Discriminator-Free Image-to-Image Translation.” arXiv, Apr. 28, 2022. doi: 10.48550/arXiv.2204.13656.

[26] R. R. Nair, T. Singh, R. Sankar, and K. Gunndu, “Multi-modal medical image fusion using LMF-GAN - A maximum parameter infusion technique,” Journal of Intelligent & Fuzzy Systems, vol. 41, no. 5, pp. 5375–5386, Jan. 2021, doi: 10.3233/JIFS-189860.

[27] M. C. Eze, L. E. Vafaei, C. T. Eze, T. Tursoy, D. U. Ozsahin, and M. T. Mustapha, “Development of a Novel Multi-Modal Contextual Fusion Model for Early Detection of Varicella Zoster Virus Skin Lesions in Human Subjects,” Processes, vol. 11, no. 8, p. 2268, Jul. 2023, doi: 10.3390/pr11082268.

[28] N. Alseelawi, H. Hazim, and H. Alrikabi, A Novel Method of Multimodal Medical Image Fusion Based on Hybrid Approach of NSCT and DTCWT, vol. 18. 2022. doi: 10.3991/ijoe.v18i03.28011.