Utilizing Deep Convolutional Neural Networks and Non-Negative Matrix Factorization for Multi-Modal Image Fusion

Dr. Nripendra Narayan Das[1], Santhakumar Govindasamy[2],
Dr. Sanjiv Rao Godla[3], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[4], Dr. E.Thenmozhi[5]
Department of Information Technology, Manipal University Jaipur, Rajasthan, India[1]
Assistant Professor, Electronics and Communication Engineering, Sri Krishna College of Technology, Coimbatore, India-641042[2]
Professor, Department of CSE (Artificial Intelligence & Machine Learning),
Aditya College of Engineering and Technology-Suraplem, Andhra Pradesh, India[3]
Professor, Faculty of Informatics and Computing, UniSZA University, Malaysia[4]
Associate Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India[5]

***Abstract*—A key element of contemporary computer vision, image fusion tries to improve the quality and interpretability of images by combining complimentary data from several image sources or modalities. This paper offers a unique method for multi-modal image fusion, combining the benefits of Deep Convolutional Neural Networks (CNNs) and Non-Negative Matrix Factorization (NMF), by using current developments in deep learning and matrix factorization techniques. Deep CNNs have shown to be remarkably effective in extracting features from images, capturing complex patterns and discriminative data. A group of deep CNNs are trained using this suggested technique on a varied dataset of multi-modal images. With the help of these networks, which extract and encode pertinent characteristics from several modalities, information-rich representations may then be combined. Concatenating, the features that were derived from the CNNs throughout the fusion process results in a fused feature representation that perfectly expresses the input modalities. The main novelty is the two-stage integration of NMF: first, breaking down the fused feature representation into non-negative basis vectors and coefficients, and then, using NMF to further extract important patterns from the fused feature maps. The non-negativity requirement in NMF guarantees the preservation of the natural structures and characteristics present in the source images, resulting in fused images that are both aesthetically pleasing and semantically intelligible. Visual examination of the merged images demonstrates the method's capacity to successfully extract important information from several modalities. The better performance and robustness of the suggested approach, which has an accuracy of roughly 99.12%, are highlighted by comparison with existing fusion approaches.**

Keywords*—Image fusion; deep convolution network; non-negative matrix factorization; multi-modal images; vector space model

I. Introduction

Image fusion has a wide range of uses in both commercial and non-industrial sectors, including security. Due to technical or optical imaging limitations, only a portion of the information may be recorded in an image using a certain type of detector or firing configuration. For example, reflecting illumination data that has intensity in a constrained range and falls within a predetermined depth-of-field is a classic example of insufficient data. By combining complimentary data gathered from many source images that were taken with various sensors or optical settings, image fusion aims to create a synthesized image [1]. Following visual assignments, including video monitoring, scene comprehension, target acknowledgment, etc., benefit from a single fusion image with greater environment representations and better perception of sight. It is challenging to efficiently and rapidly explore image on image-sharing systems due to the enormous volume of images submitted to services like Flickr and Picasa. This issue can be resolved using gathering images summarization. The goal of the image collection overview is to portray a huge, multi-modal library using only a small subset of the images and labels. The small subset shows the different elements of the initial collection, such as the attribute of interest and scene category. Image collection summarization may be employed for a variety of multimedia projects, such as automatic album building, search outcome optimization, etc. [2].

Different kinds of medical images serve an essential part in clinical diagnosis in contemporary medicine and are quite helpful in identifying disorders. Doctors typically need to integrate numerous different kinds of medical images from the same location in order to gather sufficient data for an appropriate evaluation, which frequently causes significant difficulty. When a doctor simply uses his or her own theories and conceptions to analyze a variety of medical visuals, the evaluation's objectivity is compromised and it's possible that some of the image's data is overlooked. Techniques for image fusion offer a practical solution to these problems. The collected healthcare images from various modes contain supplementary as well as duplication of data as the range of medical imaging technologies grows [3]. Other research has used a combination of verbal and graphic data to create image representations [4]. To create the visual short, the investigator developed an overview challenge that involved locating subset image examples using a homogeneous and heterogeneous message transmission technique. The image summary challenge was transformed into a hyper-graph division issue

through research, which took into account both visual and textual aspects. It suggested a max-margin assistance vector machine-based technique to extract visual ideas from multimodal dataset [5].

An imaging equipment, such as a digital single-lens reflex the device, frequently finds it challenging in the field of electronic imaging to take an image in which all the objects are sharply focused [6]. Only subjects in the depth-of-field (DOF) of an optically lens will usually look crisp in a shot at a given focal length; subjects outside the DOF will most likely to be blurry. Multi-focus image fusion, which combines many images of the same subject captured at various focal lengths to create an all-in-focus appearance, is a common approach. Additionally, a significant area within the field of image fusion is multi-focus fusion of images [7]. Many techniques for merging multi-focus images may be used, regardless of alterations, for additional image fusion applications like visible-infrared image fusion and multi-modal healthcare image synthesis. Investigating multi-focus image fusion has dual significance from this perspective, making it an explosive subject in the image computing field. Several image fusion techniques have been developed in recent years, and these techniques may be loosely divided into two distinct groups: transformation area techniques and spatially domain technique [8]. Data fusion and mining include integrating and analysing many data sources to draw out insightful conclusions and patterns. To get a complete picture of the data landscape, it aggregates data from multiple heterogeneous sources, including databases, sensors and social media. When data from many sources are combined, conflicts are resolved, and a single dataset is produced that more accurately depicts the underlying phenomenon [9]. Then, using data mining techniques, relevant relationships, trends, and patterns are identified from the pooled data. Using techniques from machine learning as well as statistical methodologies, anomalies, hidden trends, and predictions or suggestions are found in this process. Data fusion and mining are widely used in a variety of industries, such as health care, banking, marketing, and cybersecurity, and they allow companies to make data-driven decisions that improve productivity, efficiency, and decision-making [10].

To generate the multi-modal overview successively, Camargo and Gonzlez i [11] used convex non-negative matrix factorization (convex NMF) to visual modalities expressed as BoW and textual modalities expressed as vector space model (VSM). However, they did consider the sequential association between the images and labels. The characteristics of the literary topic were first taken into account. Next, images were used as inspiration for the visual concept. The sequential method, however, limits the dissemination of information from various data. They primarily relied on the textual aspects of visual summaries, ignoring the visual aspects of the literary issue and the diverse interactions between the two mediums. As a result, older summary techniques are unable to generate outcomes that exactly match the initial collection. Early spatial domain approaches frequently employed block-based fusion. Depending on the subject of the images, blocks of various sizes can be created adaptively from the images. The concept of block-based techniques is shared by a different class of spatial domain techniques that rely on image segmentation. However, the effectiveness of the classification has a big influence on how effectively these tactics work together. Many unique gradient-based, pixel-based spatial domain approaches have been created recently that can produce state-of-the-art multi-focus image fusion results. These approaches usually use rather complex fusion strategies (which can be thought of as rules in a broad sense) to their computation findings from activity degree analysis in order to boost the fusion efficacy.

The key contributions of the Multi-Modal Image Fusion using Deep Convolutional Neural Networks (CNNs) and Non-Negative Matrix Factorization (NMF) approach are:

- By concatenating the features obtained from the CNNs during the fusion process, a fused feature representation is produced that accurately captures the essence of the input modalities, improving image quality and interpretability.
- Deep CNNs are used to extract features from multi-modal images, showcasing their exceptional ability to capture intricate patterns and discriminative data, which are crucial for producing informative fused images.
- The integration of NMF in two stages is the primary innovation. In order to improve the fusion process, two steps must be taken: first, the fused feature representation must be broken down into non-negative basis vectors and coefficients; and second, NMF must be utilized to extract important patterns from the fused feature maps.
- The non-negativity condition in NMF makes sure that the fused images retain the organic shapes and traits that are present in the source images, resulting in fused images that are both aesthetically pleasing and semantically significant.

This article's remainder is organized as follows: In Section II, a summary of related research is provided. Section III presents the problem statement. The suggested approach's methodology and architecture are explained in Section IV of the article. The findings and subsequent discussion are covered in Section V. The conclusion is covered in Section VI.

II. Related Work

Although multi-model neuroimaging and gene identification technologies have advanced, attempts to integrate the two in order to investigate the virulence traits of schizophrenia (SZ) have been unsuccessful. Researchers suggest a unique approach known as grouping dense of joint non-negative matrices factorization on orthogonal domain to address this problem. The approach combines data from three models, single nucleotide polymorphism, and functional magnetic resonance imaging to identify risk genes, aberrant brain areas, and SZ-related epigenetic elements. For the purpose of eliminating unnecessary characteristics from the row of correlation matrix structures, researchers actively place diagonal constraints on the foundation matrix. Because data from genome-scanning provide extensive group information, researchers use three coefficients vectors that are densely

packed to enhance the features discovered. Our approach is tested using both the made-up and actual Mind Clinical Imaging Consortium (MCIC) datasets. Simulation results demonstrate that our approach outperforms rival tactics. GJNMFO identifies a set of risk genes, epigenetic variables, and aberrant brain functioning regions through the use of MCIC data in the study. These findings have significant economic and ecological ramifications, which science has proven [12].

For ground-based cloud recognition, deep neural networks have recently attracted a lot of attention. The entire focus of these techniques, however, is on extrapolating global features from visual input, which results in approximations for ground structures that are erroneous. The multi-evidence and multi-modal fusion network (MMFN), which is described in this article, is a unique technique for ground-based cloud identification that can increase cloud knowledge by fusing various signals in an integrated framework. By utilizing the attentive network and the main system, MMFN specifically uses a number of data points, such as global and local visual characteristics, from ground-based clouds images. Local visual qualities are gathered using the attention maps of the attentive system, which are constructed using fine-tuned salient aspects of convolutional stimulating structures. The multi-modal networking in MMFN is now studying the multi-modal properties of ground-based clouds. Researchers developed two fusion stages in MMFN to combine multi-modal features with local and global visual properties in order to fully fuse the multi-modal and multi-evidence visual qualities. The first multi-modal ground-based cloud database, or MGCD, is also made possible by study. It includes both the ground-based cloud images themselves as well as the multi-modal data that goes with each cloud image. When measured against state-of-the-art techniques, the MMFN obtains an identification performance of 88.63% on MGCD, proving its suitability for ground-based cloud recognition. The current study forbids the use extra factors, such as cloud basal height, for cloud characterization [13].

To achieve human-robot collaboration (HRC) in manufacturing processes, multimodal robot management must be intuitive and trustworthy. In earlier works, multimodal robotic control strategies were established. The technologies make it possible for human employees to control robots naturally without having to write brand-specific programming. However, because characteristics are not depicted consistently across multiple paradigms, the bulk of multimodal controlling robots' approaches are unreliable. In order to solve this problem, the research on reliable multimodal HRC production systems suggests a multimodal fusion architecture that makes use of deep learning. The proposed design consists of three modalities: verbal authority, hand gesture, and body motion. Three single-modal systems' characteristics are first trained to be retrieved, after which the characteristics are combined to swap representations. Tests show that the proposed multimodal fusion paradigm performs superior to the three unimodal models. The paper emphasizes the potential for applying the suggested multimodal fusion architecture to produce dependable HRC systems. The architectural concept paradigm wasn't made clear enough [14].

Radar electronic surveillance has new difficulties as a result of the emergence of cognitive wireless and electronic warfare; recognizing the signal generated by radar is a crucial component of this work. Research suggests a new radar signal recognition technique that uses non-negative matrix factorization network (NMFN) and ensemble learning. This system is capable of reliably recognizing radar signals under low signal-to-noise ratio conditions. Research investigates a method for extracting features based on a convolutional neural network at the beginning, which uses transfer learning as a way to address the issue of small sizes of samples. In order to extract characteristics and eliminate redundant data, research also suggests a non-negative matrix factorization system. In the third step, research create a feature fusion method utilizing stacked autoencoders (SAE), which can collect key feature expressions and condense feature dimensions. Last but not least, researcher suggests the improved artificial bee colony algorithm (IABC) as an ensemble learning technique that can increase the recognition rate. According to the simulation outcomes, recognition rates are 94.23% at 4 dB and 99.82% at 6 dB [15].

Dynamic MRI was used as a technique to record the body's various organs successive anatomy as they change over time. Nevertheless, due to mechanical and physiological limitations, its uses are restricted by shorter acquisition times. It has been demonstrated that dynamic MRI has spatio-temporal heterogeneity in its frequency spectrum (k-space). Lowering the number of k-space examples can greatly shorten the acquisition duration, yet at the expense of introducing artefacts into the associated image realm. To speed up the whole acquisition procedure, Shashidhar and Subha [16] created a cascaded Convolutional Long Short-Term Memory (ConvLSTM) framework for T2-weighted dynamic MRI patterns restoration from significantly under-sampled k-space information. Particularly, a Cartesian inadequate sampling mask could be used to under-sample completely sampled information obtained from the ADNI dataset. The aliasing artefacts caused by inadequate sampling are then eliminated using the ConvLSTM framework that has been suggested. In order to rebuild the imagery effectively and more accurately than CNN-based restoration, the ConvLSTM framework also learns the imagery's temporal and spatial connections. The utilization of medical databases presents ethical issues about data protection and informed approval, like the ADNI database. It is crucial to confirm that the research complies with ethical standards and has gotten the necessary rights and authorization for the use of the data.

III. PROBLEM STATEMENT

The requirement for efficient multi-modal image fusion to improve image quality and interpretability across many applications is the issue this research attempts to solve. Integrating data from several sources while preserving the accuracy of the original data are difficult. Complex patterns and distinguishing traits are frequently difficult to capture using traditional techniques. The work suggests a remedy that combines the capacities of CNNs and NMF to address this. Utilizing CNNs' feature extraction abilities, the idea is to produce a fused feature representation that is then improved by NMF to uncover useful patterns. Since.the non-negativity

requirement in NMF, the fused images are coherent and meaningful since their inherent structures are preserved. The suggested approach's capacity to maintain crucial diagnostic data and improve fusion quality is assessed by quantitative metrics and visual evaluations, demonstrating its potential to help precise decision-making and analysis in areas like medical imaging [17].

IV. PROPOSED FRAMEWORK

There are multiple steps in the suggested methodology for Deep Convolutional Neural Networks (CNNs) to fuse multimodal images. The process for Multi-Modal Image Fusion using Deep Convolutional Neural Networks and Non-Negative Matrix Factorization is depicted in Fig. 1. The input images 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 images and a fusion-specific loss function, 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. NMF is used to decompose extracted feature maps into non-negative basis vectors and coefficients in the setting of non-negative matrix factorization (NMF) feature extraction. The most discriminative basis vectors capturing the essential features of each modality are selected, and fusion weights are learned or fusion rules are applied to combine them. The fused basis vectors are multiplied with the corresponding coefficients to reconstruct the fused feature maps, which are then aggregated to generate the final fused image. Post-processing techniques like denoising or sharpening can be applied for further enhancement. This methodology is primarily used for multi-modal image fusion, where images from different modalities, such as infrared and visible, are combined to provide a comprehensive understanding of a scene. On the other hand, non-multi-modal image fusion is utilized in feature fusion scenarios where features from the same modality but captured under different conditions, such as exposure or focus, are fused to create a more comprehensive feature representation.

A. Data Collection

MRI brain images of 1000 datasets including healthy and unhealthy are used in the research. Among these 50% of images are used as training data and 50% of images are used as testing data. The collected brain cancer images were existing on the Kaggle depository website [18]. The datasets are distributed in Table I.

B. Data Preprocessing

Magnetic Resonance Imaging (MRI) imageries were impacted by unrelated and erratic noisy data, such as Gaussian noise and Speckle sound, which reduced the analysis value of those sample imageries. Speckle sounds have a significant impact on the contrast resolution of MRI brain imaging. Therefore, the original Hannmean filter is used to reduce noise in MRI brain images. A Hannmean filter is a filter that combines the Hanning window and Mean filters. The established Hannmean filter is used to minimize the noise in an image as well as any spatial intensity derivatives that may be present. In order to exchange each pixel's value with its surrounding neighbours' mean image values and ignore the unreliable pixel value of their image background, the Hannmean filter is used. Noises are produced in MRI brain scans by the device's inhomogeneity dis a magnetic region afforded by temperature, the malfunction of the scanner, and the patient's movement throughout the scanning process. Both noiseless methods and image resolution were used to get a crisp MRI brain image [19].

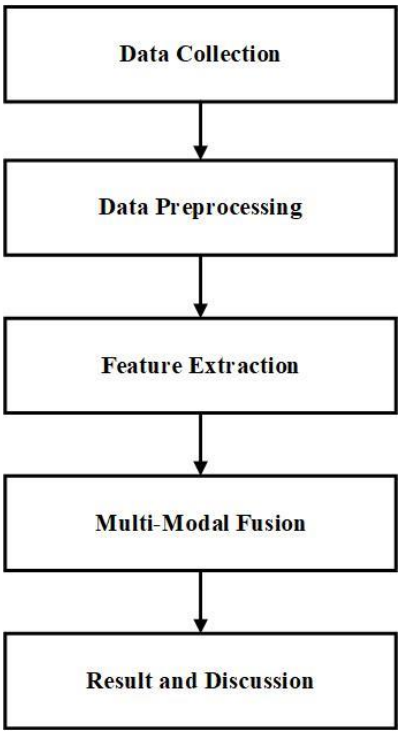

Fig. 1. Proposed framework.

TABLE I. THE COLLECTED DATASETS

	Training data	Testing data
Unhealthy	250	250
Healthy	250	250
Overall data	500	500

C. Feature Extraction using CNN

Deep CNNs are remarkably good at capturing hierarchical representations and complicated patterns. In order to begin the feature extraction process, a series of deep CNNs are trained on a variety of datasets made up of multi-modal images. These networks learn to recognize distinguishing elements that are pertinent to each input modality since they are designed to the specifics of the input modalities. CNNs extract features that capture detailed textures, forms, and structures unique to each modality by utilizing both low-level and high-level filters. The cross-modal linkages are preserved while modality-specific subtleties are captured in the learnt features. Concatenating the retrieved features from the different CNNs results in the formation of the fused feature representation, this completely embodies the essence of the multi-modal inputs. The basis for further processing, such as the usage of NMF to hone and extract more abstract patterns, is this fused feature representation. The suggested framework improves the fusion

process by utilizing CNNs' advantages in extracting valuable features, thereby assisting in the creation of high-quality fused images that capture the complimentary information available in the multi-modal data.

The suggested medical image fusion architecture contains three primary phases, which are depicted in Fig. 2. To begin, it creates the same-size weight map (m) for source images X and Y of arbitrary size using Siamese network architecture. The produced weight map X is then subjected to Gaussian gradient deconstruction to produce the matching multi-scale sub-decomposed image Gm, which is used to establish the fusion operation in the coefficient calculation merger procedure. The top layer and the remaining layers of the sub-decomposed image are represented by the symbols $G_{M,i=N}^{i,s}$ and $G_{M,0\leq1\leq N}^{i,s}$ [20]. The contrast gradient is used to break down the source images X and Y. For the following coefficient fusion technique, the multi-scale sub-decomposed images $Q_x amd\ Q_y$ are acquired. The top layers of the sub-decomposed images $Q_x amd\ Q_y$ are $Q_{X,i=N}^{i,s}$ and $Q_{Y,i=N}^{i,s}$ correspondingly. In order to denote the other layers of the sub-decomposed images$Q_x amd\ Q_y$, accordingly, research adopts the notation $Q_{X,i=N}^{i,s}$ and $Q_{Y0\leq1\leq N}^{i,s}$. Finally, distinct thresholds are established, one for the top level and the other for the layers that make up sub-decomposed image F_q.

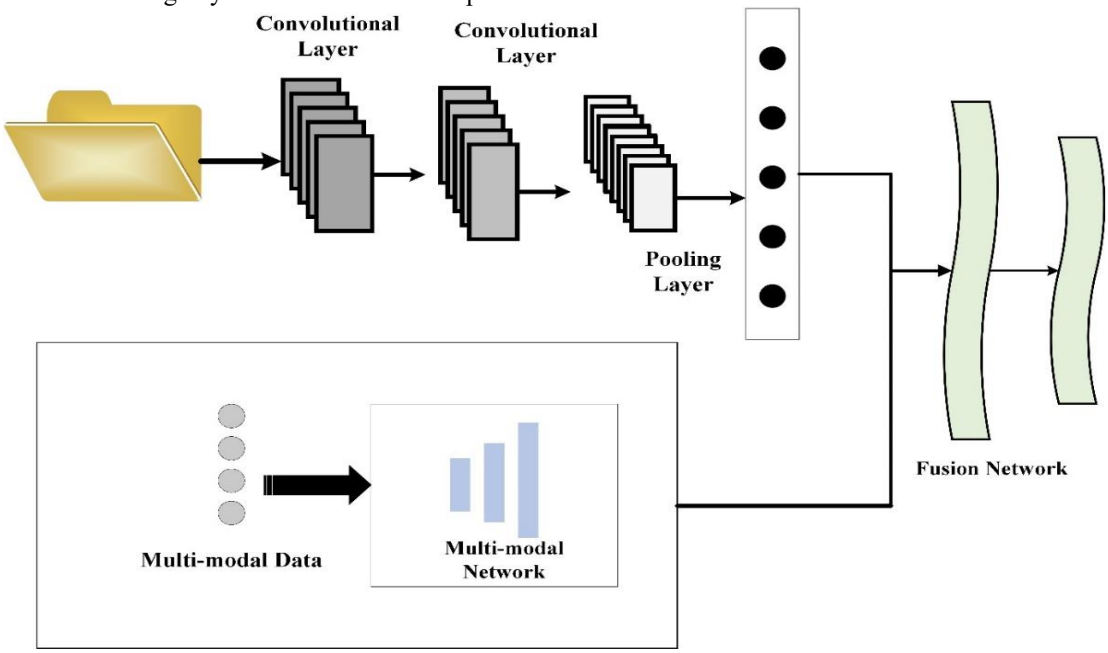

Fig. 2. Multi-modal fusion based deep convolution network.

The Fig. 2 represents a Multi-Modal Image Fusion based Deep Convolution Network, a powerful technique that combines information from different image modalities to generate a fused image with improved quality and interpretability. The network consists of input layers for each modality, followed by convolutional layers that extract relevant features from the images. Pooling layers down sample the feature maps, while fusion layers combines the extracted features from different modalities to create a comprehensive representation. Fully connected layers further transform and abstract the fused features, leading to an output layer that generates the final fused image. This architecture allows the network to leverage the strengths of each modality and enhance the understanding of the scene, making it a valuable tool in various applications [21].

The suggested technique uses CNN to accomplish an estimation of the ideal pixel level of activity and distributed weight by obtaining a weighted map of pixel activity details from numerous source images. In this study, Siamese networks are used to increase the effectiveness of CNN instruction. The Siamese system has two divisions. There are three convolutional layers and one max-pooling layer on each branch. Convolutional neural networks comprise the top two layers. The input image's non-negative matrix factorization feature extraction is done on the first layer. There are more feature maps in the second layer. The top convolutional layer extracts the characteristics of the output map. In the proposed methodology, a max-pooling layer is included as the third layer in the network architecture. This layer serves to further reduce the number of parameters and remove unnecessary samples from the feature map. By down sampling the input feature map, the max-pooling layer retains the most significant information while discarding less relevant details, effectively reducing the computational complexity. Following the max-pooling layer, a fourth layer is introduced as a convolution layer. This layer extracts more intricate and detailed information from the pooled feature map, capturing finer patterns and features. To minimize the training complexity and memory usage, a lightweight network structure is employed for this convolutional layer [22].

Specifically, the feature maps from each branch are concatenated together in the network's final stage. Concatenation gives a more thorough representation by allowing the integration of knowledge gained from several

branches. The concatenated feature maps are then immediately coupled to a two-dimensional vector using a completely connected layer. The next bi-directional SoftMax layer uses this vector as its input. The two-dimensional vector is categorized based on probability values by the bi-directional SoftMax layer, which also forecasts the probability distribution of several qualities. This forecast is essential for estimating the probability that various attributes will appear in the input data. In order to predict attributes, the network outputs a probabilistic classification by mapping the two-dimensional vector to the SoftMax layer. Overall, to extract and express complicated information from the input data, this technology combines max-pooling, convolutional layers, and concatenation of feature maps. Accurate attribute prediction based on learned representations is made possible by the fully connected layer, the bi-directional SoftMax layer, and the probability-based categorization. This method is appropriate for a variety of applications requiring attribute prediction or classification since it minimizes the number of parameters, optimizes training complexity, and increases memory efficiency [23].

This study uses a SoftMax classifier to determine the categorization probability using Eq. (1) in order to achieve categorization in the DCNN network:

$$f(r_u) = \frac{e^{r_u}}{\sum_{v=1}^{n} e^{r_u}} \quad (1)$$

The mapping between each element of one (r_u)will be approximately 1 and the rest will be closest to 0, normalizing all input matrices if one pi is greater than every other component r. The SoftMax loss curve is found as Eq. (2) when the number of batches is set to 128:

$$B = \sum_{u=0}^{size} -logf(r_u) \quad (2)$$

Stochastic gradient descent is utilized to minimize the loss function with the SoftMax loss value serving as the optimization objective. The acceleration loss and weight decay are established as the initial parameter values, respectively. Consequently, the weights are updated using Eq. (3):

$$s_{u+1} = s_u + t_u + 1 \quad (3)$$

In Eq. (3) the dynamic factor is defined as t_u and the weight is denoted as s_u at uth iteration.

D. NMF for Multi Modal Image Fusion

By permitting the breakdown of fused feature representations into non-negative basis vectors and coefficients, NMF plays a crucial role in the field of Multi-Modal Image Fusion. This decomposition technique perfectly reflects the properties of image data, where pixel values are always positive. NMF makes it easier to create a fused image while maintaining the underlying natural structures and attributes existing in the input modalities by imposing this non-negativity requirement. The basis vectors, which depict fundamental patterns shared by all modalities, identify crucial characteristics that are similar to all inputs. By permitting the breakdown of fused feature representations into non-negative basis vectors and coefficients, NMF plays a crucial role in the field of Multi-Modal Image Fusion. This decomposition technique perfectly reflects the properties of image data, where pixel values are always positive. NMF makes it easier to create a fused image while maintaining the underlying natural structures and attributes existing in the input modalities by imposing this non-negativity requirement. The basis vectors, which depict fundamental patterns shared by all modalities, identify crucial characteristics that are similar to all inputs.

Non-negative matrix factorization (NMF) is a powerful technique for multi modal fusion that aims to decompose a given data matrix into two non-negative matrices: a basis matrix and a coefficient matrix. In the context of feature extraction, NMF allows the extraction of meaningful and interpretable features by representing the input data as a linear combination of basis vectors. The basis matrix captures the fundamental components or patterns present in the data, while the coefficient matrix indicates the contribution of each basis vector to reconstruct the original data [24]. NMF assures that the extracted features are additive and non-competitive by applying non-negativity restrictions. This can be helpful for a variety of applications, including text mining, audio analysis, and image processing. The resulting basis vectors give the input data a condensed representation by emphasizing the key traits and bringing down the dimensionality, making it easier to perform further analysis or classification tasks. Overall, NMF-based feature extraction provides a practical method for identifying latent characteristics in data, enhancing the representation, comprehension, and use of complicated datasets [25].

For the assessment of non-negative matrices, the non-negative matrix factorization is used. $A \in U_{G\times U}^{+}$ and $B \in U_{U\times J}^{+}$ in which the two-matrix multiplication is similar to non-negative matrix $C \in U_{G\times U}^{+}$ could be computed using the Eq. (4):

$$C = AB + F \quad (4)$$

Where $F \in U_{G\times j}$ is an error matrix. The cost function connecting C and AB is minimized to predict the matrix of A and B as:

$$A = \arg\min_A Y\,(C|AB)\ \ for\ fixed\ B \quad (5)$$

$$B = \arg\min_B Y\,(C|AB)\ \ for\ fixed\ A \quad (6)$$

In Eqs. (5) and (6) the space between the two matrices of K and L is defined as V(K|L).

The magnitude spectrogram of the signals is frequently used as the input matrix I in various applications of non-negative matrix factorization (NMF) for acoustic signals. In this instance, the frequency content of the acoustic wave over time is represented by the matrix I. Two non-negative matrices, A and B, are created by factorizing the matrix V. The spectrum features are represented by the matrix A, where each column vector represents a particular frequency structure or spectral component. The matrix B, on the other hand, reflects the temporal activations of acoustical events. Each row vector in this matrix represents the temporal envelope of a particular event. Research has been able to roughly reconstitute the magnitude spectrogram by multiplying matrices A and B. Consider a musical signal made up of three musical events to demonstrate this idea. Each column vector in matrix A would

be able to represent a different spectral pattern or frequency structure connected to the occurrences. The temporal envelopes of the various musical events would be represented by the row vectors of the matrix B, which would show how their amplitudes changed over time.

A conversion phase is used during the image testing and fusion process on the entirely linked layer to allow processing of sources of any size. The fully connected layer is converted into two identical convolutional layers with the same kernel size. Afterwards, the network may process any size images X and Y together in order to create a dense prediction map I. Each forecast I_s on the map has a two-dimensional vector with values ranging from 0 to 1. To make the weights assigned to corresponding image blocks simpler, if one dimension of a prediction is larger than the other, it is normalized to 1 while the other dimension is set to 0. This ensures that the weight of every image block is decreased with an output dimension value of 1. In their related image blocks, two close forecasts in S have overlapped areas. The mean value of the overlapping image blocks is obtained by adding the weights of the images in these overlapped sections. With the help of this method, the network can be fed images of any size, both X and Y, and a weight map W of the same size is produced. This makes sure that each image block's weight is reduced with an output dimension value of 1. The linked image blocks of the two close forecasts in I have overlap sections. The weights of the images in these overlapped portions are added to determine the mean value of the overlapping image blocks. Using this method, the system can produce a weight map W that is the same size as an image and accept images of any size, X and Y [19].

V. Result and Discussion

Accuracy, Recall, Precision, F1-score, False Detection rate, Sensitivity, and Specificity are a few of the metrics used to verify the effectiveness of the projected model. True positive (t_p), false negative (f_n), false positive (f_p), and true negative (t_n) values are the fundamental variables that need to be computed.

A. Accuracy

It gauges how precisely the system paradigm functions. In general, it refers to the ratio of correctly observed measurements to all data. The accuracy is presented in Eq. (7) as,

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{7}$$

B. Precision

The number of right positive estimates multiplied by the total number of positive guesses is used to measure precision. It is the percentage of precisely fused multi-modal medical images. Using Eq. (8), the precision is calculated as,

$$\Pr ecision = \frac{t_p}{t_p + f_p} \tag{8}$$

C. Recall

Recall is defined as the ratio of true positives and false negatives to correct positive forecasts. It indicates the percentage of predictions that were accurate. multiple-modal image fusion. Eq. (9) is used to represent recall:

$$\text{Re}call = \frac{t_p}{t_p + f_n} \tag{9}$$

D. Sensitivity

It is a measure of the proportion of correctly foretold true positives. Eq. (10) is used to calculate sensitivity as,

$$Sensitivity = \frac{t_p}{t_p + t_n} \tag{10}$$

E. Specificity

The degree gauges how many precisely identifiable true negatives there are. Eq. (11) is used to calculate the specificity value as,

$$Specificity = \frac{t_n}{f_p + t_n} \tag{11}$$

TABLE II. Comparison of Performance Metrics

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Wavelet Transform	98.34	93.12	94.36	98.33
Fuzzy Logic	97.55	96.77	95.76	97.52
PCA	98.11	98.14	97.87	96.85
Proposed CNN-NMF	99.12	98.56	98.33	98.25

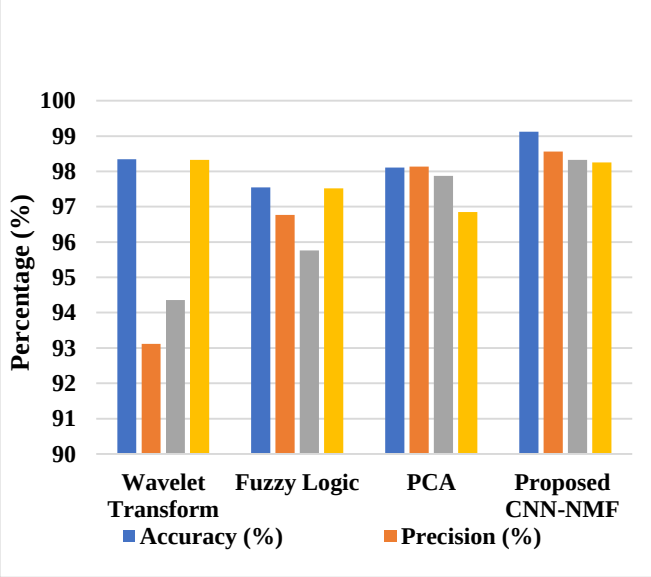

Fig. 3. Comparison of existing and proposed methods.

The Table II displays the accuracy, precision, recall, and F1-score performance evaluation of several image fusion techniques. The suggested CNN-NMF fusion strategy stands out among the tested techniques with the best accuracy of 99.12%, illustrating its capacity to successfully integrate multi-modal data. Additionally, this approach achieves impressive accuracy, recall, and F1-score values of 98.56%, 98.33%, and 98.25%, demonstrating its competence in accurately recognizing positive cases and minimizing false positives and negatives. The proposed CNN-NMF approach outperforms competing techniques like Wavelet Transform, Fuzzy Logic, and PCA, but it also has the potential to improve multi-modal image fusion tasks by capturing intricate patterns and preserving the integrity of the original data. It is depicted in Fig. 3.

TABLE III. MEDICAL IMAGE FUSION COMPARISON

Methods	Tsallis entropy	Gradient-based quality	Information ratio	Mutual information	Processing Time
MST-SR	64%	39%	36%	97%	15.05
NSCT-PC	71%	44%	40%	90%	3.77
ASR	66%	35%	39%	68%	6.15
CNN-LIU	62%	62%	28%	87%	14.58
Proposed	98%	45%	41%	92%	12.86

For the fused model with trainable and non-trainable weights, Fig. 4 and 5 displays the training accuracy and loss. Fig. 4 and 5 can be compared, and it is obvious that the model with trainable weights exhibits a faster improvement in accuracy and loss than the model with non-trainable weights. However, both networks achieve a point of convergence after around 40 epochs, with a training accuracy of about 98.07% and a loss of 0.0496. The fused model achieves a remarkable accuracy of 99.58% for the test dataset. These results show that both models eventually perform at a similar level in terms of accuracy and loss, however the model with trainable weights shows faster early development

A comparison of various methodologies based on various evaluation indicators and processing time is presented in Table III and Fig. 6. The NSCT-PC, CNN-LIU, ASR, MST-SR, and proposed algorithms are the ones that were tested. The Proposed technique receives the best score of 98% for Tsallis entropy, demonstrating its efficacy in maintaining information during the fusion process. While MST-SR, ASR, and CNN-LIU score lower with 64%, 66%, and 62% correspondingly, NSCT-PC comes in second with 71%. CNN-LIU receives the greatest score for gradient-based quality (62%), demonstrating its capacity to catch fine gradients in the fused image. The Proposed technique and ASR score 45% and 35%, respectively, whereas NSCT-PC scores 44%. The Proposed method achieves a 41% information ratio, showing a balanced preservation and utilisation of information. Following with 40% is NSCT-PC, followed by ASR with 39% and CNN-LIU with 28%. The Proposed technique receives a 92% for mutual information, demonstrating a high degree of mutual dependence between the input images in the fused result. Following with 90% is NSCT-PC, and CNN-LIU comes in at 87%. The scores for MST-SR and ASR are lower, at 97% and 68%, respectively. In terms of processing speed, NSCT-PC performs the best with a time of 3.77. The Proposed approach achieves 12.86, whereas ASR comes in second with 6.15. The processing times for MST-SR and CNN-LIU are 15.05 and 14.58, respectively. The Proposed method achieves competitive scores for gradient-based quality and stands out in terms of Tsallis entropy, information ratio, and mutual information. A promising method for multi-modal image fusion, it surpasses competing techniques in most assessment measures despite taking a little longer to process data than NSCT-PC.

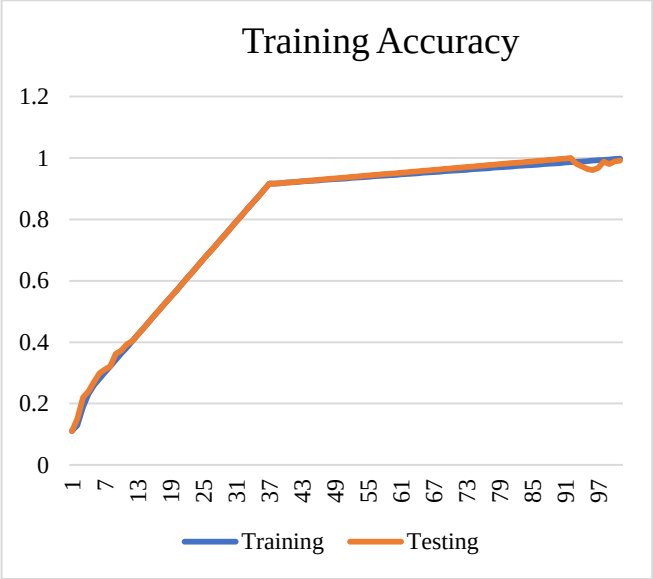

Fig. 4. Training accuracy.

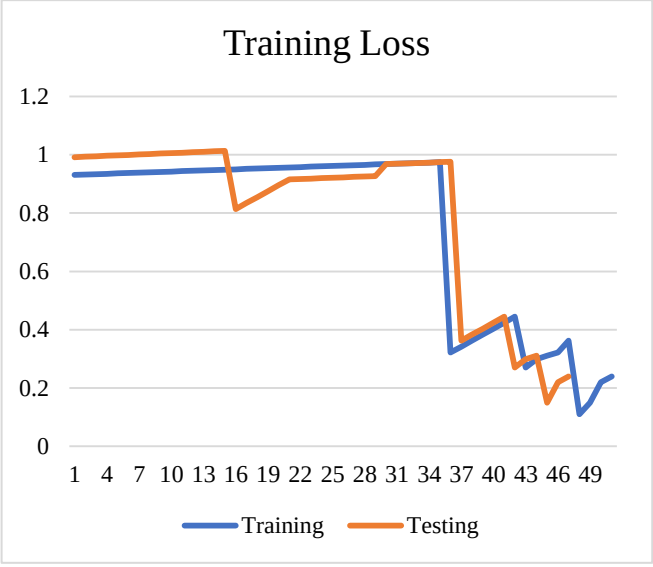

Fig. 5. Training loss.

F. Discussion

The proposed research offers a novel method for multi-modal image fusion within the context of modern computer

vision that makes use of both Deep CNNs and NMF's advantages. Deep CNNs have been shown to be adept at extracting minute details and identifying subtle patterns in images, making them an invaluable tool for working with multi-modal data. The paper suggests using these deep CNNs in a two-stage fusion process. First, the neural networks are trained to extract significant features from various modalities, and then the features that were extracted are concatenated to provide a thorough fused picture of the input data. This method stands out due to the creative ways in which NMF is applied in two different phases: first, to break down the fused representations of features into non-negative basic vector and coefficients, and subsequently, to further extract significant patterns from the resulting fused feature maps. The inherent non-negativity requirement in NMF guarantees the preservation of organic structures and inherent qualities in the source images, producing fused images that are visually beautiful and semantically comprehensible. The method excels at extracting critical information from many modalities, as shown by a visual analysis of the fused images. Its amazing accuracy also stands out as a noteworthy accomplishment, beating other fusion techniques and demonstrating its better performance and resilience. As a result of the partnership between deep CNNs and NMF, this work offers an appealing method for multi-modal picture fusion that yields a reliable and highly precise fusion technique. The suggested method successfully collects and combines data from several modalities, producing combined images that are not only aesthetically pleasing but also semantically relevant. This is made possible by successfully merging both cutting-edge deep learning methods with matrix factorization techniques. This multi-modal fusion invention is poised to make major strides in a number of sectors that depend on picture processing and interpretation and where it is crucial to accurately extract complementing information from many sources.

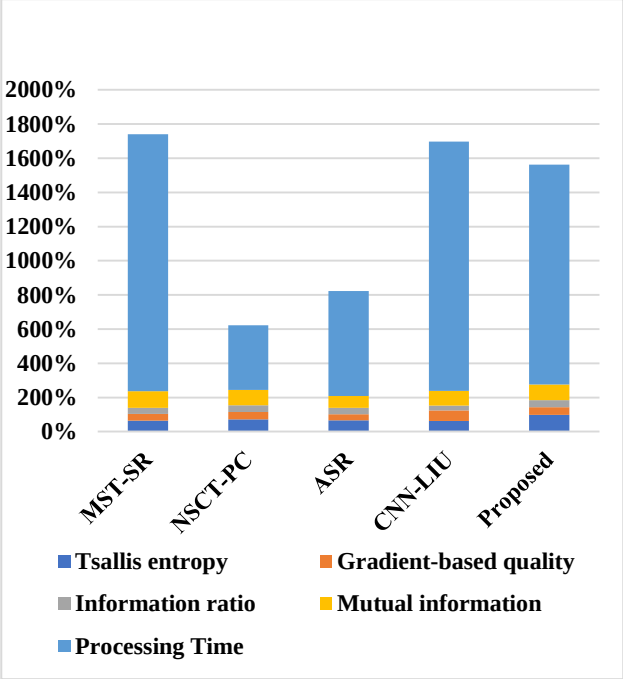

Fig. 6. Objective evaluation comparison.

VI. Conclusion

In this research, a unique and efficient multi-modal image fusion approach that makes use of Deep CNNs and NMF is provided. The suggested method tackles the fundamental problem of improving image quality and interpretability through fusion by taking use of current developments in deep learning and matrix factorization techniques. Deep CNNs have been shown to be effective in extracting features from a variety of input modalities, underscoring its importance in this situation by capturing complex patterns and discriminative data necessary for successful fusion. The approach creates information-rich representations that are then smoothly merged via the fusion process by training a series of deep CNNs on a variety of datasets. By allowing the extraction of crucial patterns from fused feature representations while conserving the inherent structures of the source images, the dual-stage integration of NMF represents a singular invention. This preservation, which is grounded in NMF's non-negativity condition, produces fused images that are both aesthetically cohesive and semantically understandable. The visual proof of information effectively collected from many modalities supports the approach's potential even more. This study's overall findings represent a substantial improvement in multi-modal image fusion, with potential applications in industries that need precise data integration and image enhancement.

VII. References

[1] J. Gao, Y. Lu, J. Qi, and L. Shen, “A radar signal recognition system based on non-negative matrix factorization network and improved artificial bee colony algorithm,” IEEE Access, vol. 7, pp. 117612–117626, 2019.

[2] F. Behrad and M. Saniee Abadeh, “An overview of deep learning methods for multimodal medical data mining,” Expert Syst. Appl., vol. 200, p. 117006, Aug. 2022, doi: 10.1016/j.eswa.2022.117006.

[3] S. Zheng, B. Fang, L. Li, M. Gao, Y. Wang, and K. Peng, “Automatic liver lesion segmentation in CT combining fully convolutional networks and non-negative matrix factorization,” in Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound: International Workshops, BIVPCS 2017 and POCUS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings, Springer, 2017, pp. 44–51.

[4] K. C. Ravikumar, P. Chiranjeevi, N. Manikanda Devarajan, C. Kaur, and A. I. Taloba, “Challenges in internet of things towards the security using deep learning techniques,” Meas. Sens., vol. 24, p. 100473, Dec. 2022, doi: 10.1016/j.measen.2022.100473.

[5] A. Khalil, M. Elmogy, M. Ghazal, C. Burns, and A. El-Baz, “Chronic wound healing assessment system based on different features modalities and non-negative matrix factorization (nmf) feature reduction,” IEEE Access, vol. 7, pp. 80110–80121, 2019.

[6] B. Lin, X. Tao, and J. Lu, “Hyperspectral image denoising via matrix factorization and deep prior regularization,” IEEE Trans. Image Process., vol. 29, pp. 565–578, 2019.

[7] B. Swiderski, J. Kurek, S. Osowski, M. Kruk, and W. Barhoumi, “Deep learning and non-negative matrix factorization in recognition of mammograms,” in Eighth International Conference on Graphic and Image Processing (ICGIP 2016), SPIE, 2017, pp. 53–59.

[8] H. Xu and J. Ma, “EMFusion: An unsupervised enhanced medical image fusion network,” Inf. Fusion, vol. 76, pp. 177–186, 2021.

[9] B. Lin, X. Tao, and J. Lu, “Hyperspectral Image Denoising via Matrix Factorization and Deep Prior Regularization,” IEEE Trans. Image Process., vol. 29, pp. 565–578, 2020, doi: 10.1109/TIP.2019.2928627.

[10] S. Mirzaei, S. Khosravani, and others, “Hyperspectral image classification using non-negative tensor factorization and 3D convolutional neural networks,” Signal Process. Image Commun., vol. 76, pp. 178–185, 2019.

[11] D. Li, Z. Gao, X.-P. Zhang, G. Zhai, and X. Yang, "Generative adversarial networks for non-negative matrix factorization in temporal psycho-visual modulation," Digit. Signal Process., vol. 100, p. 102681, 2020.

[12] P. Peng et al., "Group sparse joint non-negative matrix factorization on orthogonal subspace for multi-modal imaging genetics data analysis," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 19, no. 1, pp. 479–490, 2020.

[13] S. Liu, M. Li, Z. Zhang, B. Xiao, and T. S. Durrani, "Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition," Remote Sens., vol. 12, no. 3, p. 464, Feb. 2020, doi: 10.3390/rs12030464.

[14] H. Liu, T. Fang, T. Zhou, and L. Wang, "Towards Robust Human-Robot Collaborative Manufacturing: Multimodal Fusion," IEEE Access, vol. 6, pp. 74762–74771, 2018, doi: 10.1109/ACCESS.2018.2884793.

[15] J. Gao, Y. Lu, J. Qi, and L. Shen, "A Radar Signal Recognition System Based on Non-Negative Matrix Factorization Network and Improved Artificial Bee Colony Algorithm," IEEE Access, vol. 7, pp. 117612–117626, 2019, doi: 10.1109/ACCESS.2019.2936669.

[16] S. V. Yakkundi and D. P. Subha, "Convolutional LSTM: A Deep learning approach for Dynamic MRI Reconstruction," in 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India: IEEE, Jun. 2020, pp. 1011–1015. doi: 10.1109/ICOEI48184.2020.9142982.

[17] W. Ma et al., "Infrared and Visible Image Fusion Technology and Application: A Review," Sensors, vol. 23, no. 2, p. 599, 2023.

[18] H. R. Almadhoun and S. S. A. Naser, "Detection of Brain Tumor Using Deep Learning," vol. 6, no. 3, p. 19, 2022.

[19] Z. Huang, "Integrative Analysis of Multimodal Biomedical Data with Machine Learning," PhD Thesis, Purdue University Graduate School, 2021.

[20] F. Gao, X. Deng, M. Xu, J. Xu, and P. L. Dragotti, "Multi-modal convolutional dictionary learning," IEEE Trans. Image Process., vol. 31, pp. 1325–1339, 2022.

[21] R. Soroush and Y. Baleghi, "NIR/RGB image fusion for scene classification using deep neural networks," Vis. Comput., vol. 39, no. 7, pp. 2725–2739, Jul. 2023, doi: 10.1007/s00371-022-02488-0.

[22] A. Khader, J. Yang, and L. Xiao, "NMF-DuNet: Nonnegative Matrix Factorization Inspired Deep Unrolling Networks for Hyperspectral and Multispectral Image Fusion," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 15, pp. 5704–5720, 2022, doi: 10.1109/JSTARS.2022.3189551.

[23] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," Comput. Biol. Med., vol. 128, p. 104129, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104129.

[24] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 502–518, 2020.

[25] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," Inf. Fusion, vol. 69, pp. 40–51, 2021.